

UNIVERSIDAD AUTÓNOMA DE MADRID  
ESCUELA POLITÉCNICA SUPERIOR



# ON THE FUSION OF SINGLE-TARGET VIDEO OBJECTS TRACKING ALGORITHMS

Rafael Martín Nieto  
Supervisor: José María Martínez Sánchez

-TRABAJO FIN DE MASTER-

Departamento de Tecnología Electrónica y de las Comunicaciones  
Escuela Politécnica Superior  
Universidad Autónoma de Madrid  
Septiembre 2013



# ON THE FUSION OF SINGLE-TARGET VIDEO OBJECTS TRACKING ALGORITHMS

**Rafael Martín Nieto**

**Supervisor: José María Martínez Sánchez**

email: {Rafael.MartinN, JoseM.Martinez}@uam.es



**Video Processing and Understanding Lab**  
**Departamento de Tecnología Electrónica y de las Comunicaciones**  
**Escuela Politécnica Superior**  
**Universidad Autónoma de Madrid**  
**Septiembre 2013**

Trabajo parcialmente financiado por el gobierno español bajo el proyecto  
TEC2011-25995 (EventVideo)







# Abstract

The final objective of this work is to create a fusion system which improves the performance of several object trackers, within a methodological and rigorous evaluation framework. The considered algorithms are monocular single target trackers.

After analyzing in detail the state of the art, an evaluation framework is selected and presented. The sequences selected in this evaluation try to consider the main problems that are faced by trackers (scale changes, illumination changes, occlusion, noise, . . . ). Then, classical and modern tracking algorithms are selected and evaluated individually, in order to understand its functioning in different scenarios and problems. Finally, some fusion methods are described and evaluated.



# Acknowledgements

En primer lugar a Chema por su interés y ayuda en mi formación, no solo durante este trabajo, si no desde el primer año de carrera.

También a todos los miembros del VPULab, con los que he pasado mucho tiempo este último año. Este laboratorio es mucho más que un grupo de trabajo.

Agradecer también a mis padres, hermana y familia, por estar siempre a mi lado y apoyarme. Las comidas familiares de los fines de semana son momentos únicos que nunca olvidaré.

Finalmente a mis amigos. Una persona no es completa sin unos buenos amigos que te acompañen a lo largo del camino.

Rafael Martín Nieto  
September 2013



# Contents

<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Objectives . . . . .	1
1.2 Document Structure . . . . .	2
<b>2 State Of The Art</b>	<b>3</b>
2.1 Introduction . . . . .	3
2.2 Object Tracking . . . . .	3
2.3 Individual Tracking Algorithms . . . . .	6
2.3.1 Template Matching (TM) . . . . .	7
2.3.2 Mean-Shift (MS) . . . . .	7
2.3.3 Particle Filter-based Colour tracking (PFC) . . . . .	8
2.3.4 Lucas-Kanade tracking (LK) . . . . .	9
2.3.5 Incremental learning for robust Visual Tracking (IVT) . . . . .	9
2.3.6 Tracking Learning Detection tracking (TLD) . . . . .	10
2.3.7 Corrected Background-Weighted Histogram tracker (CBWH) . . . . .	10
2.3.8 Scale and Orientation Adaptive Mean-Shift Tracking (SOAMST) . . . . .	11
2.4 Fusion . . . . .	11
2.4.1 Fusion architectures . . . . .	12
2.4.2 Fusion levels . . . . .	12
2.4.3 Combination techniques . . . . .	12
2.5 Conclusions . . . . .	14
<b>3 Evaluation Framework</b>	<b>17</b>
3.1 Introduction . . . . .	17
3.2 Single Object Video Tracking dataset - SOVTds . . . . .	17
3.3 Selection of evaluation metric . . . . .	21
3.3.1 Metrics . . . . .	21
3.3.2 Metrics correlation study . . . . .	25
3.4 Conclusions . . . . .	26

<b>4</b>	<b>Individual trackers evaluation</b>	<b>27</b>
4.1	Introduction . . . . .	27
4.2	Individual Tracking Algorithms . . . . .	27
4.2.1	Template matching (TM) . . . . .	27
4.2.2	Mean-Shift (MS) . . . . .	28
4.2.3	Particle Filter-based Colour tracking (PFC) . . . . .	29
4.2.4	Lucas-Kanade tracking (LK) . . . . .	29
4.2.5	Incremental learning for robust visual tracking (IVT) . . . . .	30
4.2.6	Tracking learning detection tracking (TLD) . . . . .	31
4.2.7	Corrected Background-Weighted Histogram tracker (CBWH) . . . . .	32
4.2.8	Scale and Orientation Adaptive Mean-Shift Tracking (SOAMST) . . . . .	33
4.3	Comparative results . . . . .	33
4.4	Conclusions . . . . .	36
<b>5</b>	<b>Fusion</b>	<b>37</b>
5.1	Introduction . . . . .	37
5.2	Fusion Methods . . . . .	37
5.2.1	Mean . . . . .	38
5.2.2	Median . . . . .	38
5.2.3	Majority voting . . . . .	39
5.3	Fusion results . . . . .	39
5.4	Comparative results between individual trackers and fusions . . . . .	42
5.5	Results extracted from the the SoA fusion algorithms . . . . .	45
5.6	Conclusions . . . . .	46
<b>6</b>	<b>Conclusions and future work</b>	<b>47</b>
6.1	Introduction . . . . .	47
6.2	Conclusions . . . . .	47
6.3	Future Work . . . . .	48
	<b>Bibliography</b>	<b>49</b>
	<b>List of Abbreviations</b>	<b>53</b>
	<b>Appendix</b>	<b>55</b>
<b>A</b>	<b>Video object tracking datasets</b>	<b>55</b>
A.1	SPEVI . . . . .	55
A.1.1	Single Face Dataset . . . . .	55
A.1.2	Multiple Face Dataset . . . . .	55
A.2	ETISEO . . . . .	56
A.3	PETS . . . . .	57
A.3.1	PETS2000 . . . . .	57
A.3.2	PETS 2001 . . . . .	57
A.3.3	PETS 2006 . . . . .	58

A.3.4	PETS 2007 . . . . .	58
A.3.5	PETS 2010 . . . . .	59
A.4	CAVIAR . . . . .	59
A.5	VISOR . . . . .	60
A.6	iLids . . . . .	61
A.7	Clemson dataset . . . . .	62
A.8	MIT Traffic Dataset . . . . .	63
<b>B</b>	<b>Trackers and fusion results: Data tables</b>	<b>65</b>
<b>C</b>	<b>Trackers and fusion results: Bar figures</b>	<b>73</b>
C.1	Sequence Frame Detection Accuracy (SFDA) . . . . .	73
C.2	Average Tracking Accuracy (ATA) . . . . .	75
C.3	Average Tracking Error (ATE) . . . . .	76
C.4	Overlap . . . . .	77
C.5	Area Under the lost track ratio Curve (AUC) . . . . .	79
C.6	Closeness of Track (CT) . . . . .	80
C.6.1	The closeness of the whole sequence (CTM) . . . . .	80
C.6.2	weighted standard deviation of track closeness (CTD) . . . . .	81
C.7	Track Completeness (TC) . . . . .	83
C.8	Combined Tracking Performance Score (CoTPS) . . . . .	84
<b>D</b>	<b>Trackers and fusion results: Comparative tables</b>	<b>87</b>
D.1	Individual and fusions global scores . . . . .	87
D.2	Difference global scores . . . . .	88
D.3	Percentual difference global scores . . . . .	91
D.4	Difference and percentual difference global scores compared with the best individual tracker . . . . .	95

# List of Figures

2.1	Video object tracker canonical system . . . . .	5
3.1	Sample frames of the SOVTds . . . . .	20
4.1	SFDA result for TM tracker . . . . .	27
4.2	SFDA result for MS tracker . . . . .	28
4.3	SFDA result for PFC tracker . . . . .	29
4.4	SFDA result for LK tracker . . . . .	29
4.5	SFDA result for IVT tracker . . . . .	30
4.6	SFDA result for TLD tracker . . . . .	31
4.7	SFDA result for CBWH tracker . . . . .	32
4.8	SFDA result for SOAMST tracker . . . . .	33
4.9	L1 SFDA result for all the individual trackers . . . . .	34
4.10	L2 SFDA result for all the individual trackers . . . . .	34
4.11	L3 SFDA result for all the individual trackers . . . . .	34
4.12	L4 SFDA result for all the individual trackers . . . . .	35
5.1	Fusion Block diagram . . . . .	38
5.2	Fusion SFDA result of L1 . . . . .	39
5.3	Fusion SFDA result of L2 . . . . .	40
5.4	Fusion SFDA result of L3 . . . . .	41
5.5	Fusion SFDA result of L4 . . . . .	41
A.1	Sample frames for the SPEVI dataset . . . . .	56
A.2	Sample frames for the ETISEO dataset . . . . .	57
A.3	Sample frames for the PETS2006 dataset . . . . .	58
A.4	Sample frames for the PETS2007 dataset . . . . .	59
A.5	Sample frames for the PETS2010 dataset . . . . .	59
A.6	Sample frames for the CAVIAR dataset . . . . .	60
A.7	Sample frames for the VISOR dataset . . . . .	61
A.8	Sample frames for the i-LIDS dataset . . . . .	62
A.9	Sample frames for the CLEMSON dataset . . . . .	62
A.10	Sample frames for the MIT traffic dataset . . . . .	63
C.1	Fusion SFDA result of L1 . . . . .	73
C.2	Fusion SFDA result of L2 . . . . .	74



C.3 Fusion SFDA result of L3 . . . . .	74
C.4 Fusion SFDA result of L4 . . . . .	74
C.5 Fusion ATA result of L1 . . . . .	75
C.6 Fusion ATA result of L2 . . . . .	75
C.7 Fusion ATA result of L3 . . . . .	75
C.8 Fusion ATA result of L4 . . . . .	76
C.9 Fusion ATEinv result of L1 . . . . .	76
C.10 Fusion ATEinv result of L2 . . . . .	76
C.11 Fusion ATEinv result of L3 . . . . .	77
C.12 Fusion ATEinv result of L4 . . . . .	77
C.13 Fusion PixelOverlap result of L1 . . . . .	77
C.14 Fusion PixelOverlap result of L2 . . . . .	78
C.15 Fusion PixelOverlap result of L3 . . . . .	78
C.16 Fusion PixelOverlap result of L4 . . . . .	78
C.17 Fusion AUCinv result of L1 . . . . .	79
C.18 Fusion AUCinv result of L2 . . . . .	79
C.19 Fusion AUCinv result of L3 . . . . .	79
C.20 Fusion AUCinv result of L4 . . . . .	80
C.21 Fusion CTM result of L1 . . . . .	80
C.22 Fusion CTM result of L2 . . . . .	80
C.23 Fusion CTM result of L3 . . . . .	81
C.24 Fusion CTM result of L4 . . . . .	81
C.25 Fusion CTD result of L1 . . . . .	81
C.26 Fusion CTD result of L2 . . . . .	82
C.27 Fusion CTD result of L3 . . . . .	82
C.28 Fusion CTD result of L4 . . . . .	82
C.29 Fusion TC result of L1 . . . . .	83
C.30 Fusion TC result of L2 . . . . .	83
C.31 Fusion TC result of L3 . . . . .	83
C.32 Fusion TC result of L4 . . . . .	84
C.33 Fusion CoTPSinv result of L1 . . . . .	84
C.34 Fusion CoTPSinv result of L2 . . . . .	84
C.35 Fusion CoTPSinv result of L3 . . . . .	85
C.36 Fusion CoTPSinv result of L4 . . . . .	85

# List of Tables

3.1	Complexity factors for the video tracking dataset . . . . .	19
3.2	Correlation between metrics . . . . .	26
5.1	global SFDA scores . . . . .	42
5.2	Difference (percentaje) SFDA global score between fusion trackers and individual algorithms trackers . . . . .	43
5.3	Percentual difference (percentage) SFDA global score between fusion trackers and individual algorithms trackers . . . . .	43
5.4	Difference (percentaje) global score between fusion trackers and best individual algorithms trackers . . . . .	44
5.5	Percentual difference (percentaje) global score between fusion trackers and best individual algorithms trackers . . . . .	44
B.1	Results for TM tracker . . . . .	66
B.2	Results for MS tracker . . . . .	66
B.3	Results for PFC tracker . . . . .	66
B.4	Results for LK tracker . . . . .	67
B.5	Results for IVT tracker . . . . .	67
B.6	Results for TLD tracker . . . . .	67
B.7	Results for CBWH tracker . . . . .	68
B.8	Results for SOAMST tracker . . . . .	68
B.9	Results for $\geq 1$ (OR) fusion . . . . .	68
B.10	Results for $\geq 2$ fusion . . . . .	69
B.11	Results for $\geq 3$ fusion . . . . .	69
B.12	Results for $\geq 4$ fusion . . . . .	69
B.13	Results for $\geq 5$ fusion . . . . .	70
B.14	Results for $\geq 6$ fusion . . . . .	70
B.15	Results for $\geq 7$ fusion . . . . .	70
B.16	Results for $\geq 8$ fusion . . . . .	71
B.17	Results for $\geq 9$ fusion . . . . .	71
B.18	Results for $\geq 10$ fusion . . . . .	71
D.1	Individual global scores . . . . .	87
D.2	Fusion global scores . . . . .	88

D.3	Difference (percentaje) SFDA global score between fusion trackers and individual algorithms trackers . . . . .	88
D.4	Difference (percentaje) ATA global score between fusion trackers and individual algorithms trackers . . . . .	89
D.5	Difference (percentaje) ATEinv global score between fusion trackers and individual algorithms trackers . . . . .	89
D.6	Difference (percentaje) AUCinv global score between fusion trackers and individual algorithms trackers . . . . .	89
D.7	Difference (percentaje) PixelOV global score between fusion trackers and individual algorithms trackers . . . . .	90
D.8	Difference (percentaje) CTM global score between fusion trackers and individual algorithms trackers . . . . .	90
D.9	Difference (percentaje) CTD global score between fusion trackers and individual algorithms trackers . . . . .	90
D.10	Difference (percentaje) TC global score between fusion trackers and individual algorithms trackers . . . . .	91
D.11	Difference (percentaje) CoTPS global score between fusion trackers and individual algorithms trackers . . . . .	91
D.12	Percentual difference (percentage) SFDA global score between fusion trackers and individual algorithms trackers . . . . .	92
D.13	Percentual difference (percentage) ATA global score between fusion trackers and individual algorithms trackers . . . . .	92
D.14	Percentual difference (percentage) ATEinv global score between fusion trackers and individual algorithms trackers . . . . .	92
D.15	Percentual difference (percentage) AUCinv global score between fusion trackers and individual algorithms trackers . . . . .	93
D.16	Percentual difference (percentage) PixelOV global score between fusion trackers and individual algorithms trackers . . . . .	93
D.17	Percentual difference (percentage) CTM global score between fusion trackers and individual algorithms trackers . . . . .	93
D.18	Percentual difference (percentage) CTD global score between fusion trackers and individual algorithms trackers . . . . .	94
D.19	Percentual difference (percentage) TC global score between fusion trackers and individual algorithms trackers . . . . .	94
D.20	Percentual difference (percentage) CoTPS global score between fusion trackers and individual algorithms trackers . . . . .	94
D.21	Difference (percentaje) global score between fusion trackers and best individual algorithms trackers . . . . .	95
D.22	Percentual difference (percentage) global score between fusion trackers and best individual algorithms trackers . . . . .	95



# Chapter 1

## Introduction

Computer vision is an important research field that includes methods for processing images and sequences from the real world in order to understand its content. There are many algorithms for object tracking in the state of the art, but none of the algorithms works correctly in all situations. Many of them, due to their design, function properly only for specific cases. Furthermore, there is no common evaluation framework, so most of the authors usually evaluate their algorithms to their own criteria.

Taking into account the current situation with respect to object tracking, there are two main motivations of this Master Thesis: The first one is to approach the evaluation of video object trackers in a methodological way, as it is a main aspect to evaluate and improve the results; the second one is the study of how to improve object tracking results.

The evaluation in the state of the art is performed in a relatively individualistic way, even though there are efforts trying to unify it[1]. The results are good only in specific cases. As each tracker works well depending on the scenario and problem, the fusion of multiple tracking algorithms can help to solve this problem.

### 1.1 Objectives

The final objective of this work is to create a fusion system which improves the performance of several single target object trackers, within a methodological and rigorous evaluation framework. Therefore, the work is divided in three main objectives. Firstly, an evaluation framework will be proposed. Secondly, some individual video object tracking algorithms, extracted from the state of the art, will be evaluated within the previously mentioned framework. Finally, some fusion methods will be tested with the aim of improving the individual tracker results.

## 1.2 Document Structure

The structure of the document is as follows:

- Chapter 1. This chapter introduces the work and presents the motivation and the objectives of the Master Thesis.
- Chapter 2. This chapter presents an overview of the literature related to the work presented in this Master Thesis.
- Chapter 3. This chapter presents both the used content set and the considered metrics for the evaluation of each tracker.
- Chapter 4. This chapter presents an individual evaluation of each one of the basic trackers.
- Chapter 5. This chapter presents the different fusion methods and their evaluation results.
- Chapter 6. This chapter summarizes the main achievements of the work, discusses the obtained results and provides suggestions for future work.

At the end, four appendices list further details. Appendix A presents some public available state of the art tracking datasets. Appendices B, C and D present the individual trackers and fusion methods results, for each one of the metrics: appendix B presents tables with the obtained scores, appendix C presents bar figures to facilitate comparison between different trackers and fusions, and finally appendix D shows comparative tables between the scores obtained for the individual trackers and for the fusion methods.

## Chapter 2

# State Of The Art

### 2.1 Introduction

This chapter gives an overview of previous work that has been done in the scope of the study presented in this Master Thesis. In the next sections, we describe the areas of object tracking (section 2.2), tracking algorithms(section 2.3) and fusion (section 2.4).

### 2.2 Object Tracking

Computer Vision is a field whose goal is to automate the processing of images to understand its content. Computer Vision tries to imitate the human vision system in which the brain processes images captured by the eyes. The data may have different formats such as video sequences, different views from multiple cameras or multi-dimensional data provided by medical scanners. This information is used to solve specific tasks or to understand what happens in the scene. Object tracking is one of the most important tasks in computer vision.

Video object tracking is the process of locating (or estimating) one or more moving objects of interest over time using sequences acquired by one or more cameras. In [2], the problem is defined as the task of following one or more objects in a scene, from their first appearance to their exit. In its simplest form, tracking can be defined[3] as the problem of estimating the trajectory of an object in the image plane as it moves around a scene. A tracker assigns labels to the tracked objects in the different frames of a video. Since the amount of data to be processed is very large, object tracking is a task of great complexity and great time consumption.

An object may be anything of interest within the scene that can be detected, and

depends on the requirements of the application. In a real tracking situation, both background and tracked object(s) are allowed to vary, what difficulties the tracking task. A set of constraints can be put to make this problem solvable. The more the constraints, the problem is easier to solve. Some of the constraints that are generally imposed during object tracking are[2]:

- Object motion is smooth with no abrupt changes
- There are no sudden changes in the background
- Changes in the appearance of the object are gradual
- Fixed camera scenarios
- Limited number and size of objects
- Limited amount of occlusions

Video tracking algorithms basically aim at identifying the candidate position (pixel coordinates) within a video frame where a target model is most likely to be present. The tracker objective is to model the relation between the tracked object and its corresponding pixel values of a set of frames. The target is usually a predefined region of interest (e.g., rectangle, circle, ellipse) either automatically or semi-automatically delimited within a given image.

A video tracker can be decomposed in five main logical components[4]:

- Feature extraction: The definition of a method to extract relevant information from an image area where the target object is placed. This method can be based on colour, gradient, motion, etc.
- Target representation: The definition of a representation for encoding the appearance of a target, defining the object characteristics.
- Localisation: The definition of a method to propagate the state of the target over time. The information used in this step is extracted from the two previous logical components.
- Track management: The definition of a strategy to take into account target appearing and disappearing from the image plane.
- Meta-data extraction: The extraction of meta-data (e.g. video annotation, scene understanding and behaviour recognition) to be used by the video application.



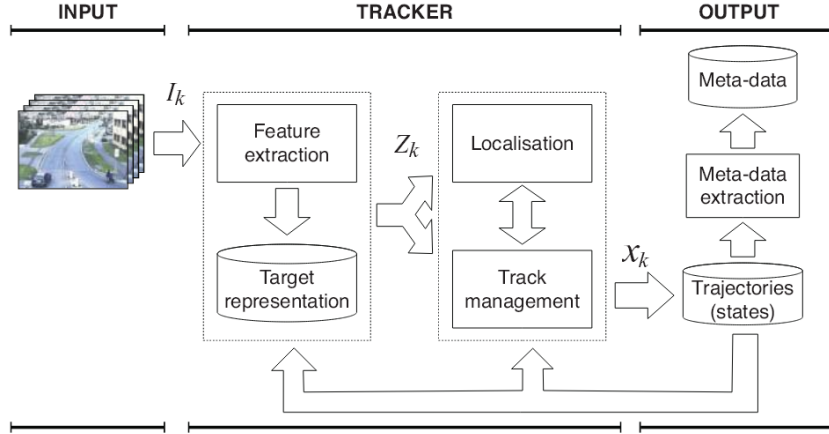


Figure 2.1: Video object tracker canonical system[4]

These logical components form the canonical system of video object trackers. Figure 2.1, extracted from [4], shows this canonical system.

Multi-tracking algorithms[5][6][7][8] have the same underlying principles as described above, although they support several simultaneous targets within the same video sequence. Thanks to the rise of surveillance systems and the increase in their complexity, there has been an increased interest in the problem of multi-target tracking[4][9][10]. The main difference between mono-target and multi-target tracking lies in the state-space model used. In a single target tracking algorithm, the state of only one target is modelled, and detections from other targets are assumed to be false alarms. Multiple target tracking algorithms consider simultaneously the existence of more than one target in its association process.

The goal of a multicamera tracking system [11][12][13][14] is to establish the correspondence between observations of objects through different cameras. To estimate the trajectory of moving objects from one place to another in the camera network, the cameras share different types of data such as objective measures (position, speed, size, ...), the target state, the estimated uncertainty (covariance matrices) and other derived measures. To effectively utilize this information, each camera must consider the others. Such environments can solve some of the problems encountered in monocular environments, such as occlusions. Nevertheless, the use of multicamera systems presents new problems as obtaining the correct association of detections between cameras or the removal of redundant information. The multicamera trackers can be classified into three main groups [15], depending on the communication between the sensors: centralized, decentralized and distributed.

The centralized tracking is performed on a single node that receives data from each

camera in the network. Although the centralized approaches can be used directly from camera trackers in the fused data, the presence of a single global fusion center often cause high data transfer rates and poor scalability and energy efficiency.

In decentralized tracking, the cameras are grouped into clusters and each node (e.g., camera) communicates with their local fusion centers. The communication overhead is reduced by limiting the communication within each cluster and among fusion centers. The characteristics of the objects are extracted in each camera view, and then are sent to the multi-tracking local fusion center. Finally, fusion centers communicate with each other through the network. The fusion centers are network nodes that collect data from the cameras within a cluster. The use of fusion centers favors scalability and reduces total communication load.

To further increase the scalability and to reduce the cost of communication, the distributed tracking functions without local fusion centers. The estimations generated in a camera are transmitted only to its immediate neighbors. The estimations received are used to refine the estimations of the next camera and these estimations are refined and transmitted to the rest of its neighbors. This process is completed after a predefined number of steps, after visiting all cameras, or when the uncertainty has decreased below a desired value.

In general, other sensors can be taken into account to help the analysis of the video with additional information (depth, audio location, etc.). These sensors allow to use information that can not be achieved with conventional video cameras, what can improve the performance of the analysis. Despite this, the new sensors may have additional restrictions that must be taken into account, such as maximum distance of sensitivity exhibited by some depth sensors.

In practice, the results of the different stages of analysis are interconnected (for example, an erroneous foreground/background segmentation significantly complicates tracking objects). This has led to the development of techniques using the results of the steps of high-level analysis to improve the results obtained in low-level stages[16].

## 2.3 Individual Tracking Algorithms<sup>1</sup>

There are multiple tracking algorithms in the state of the art. The following subsections describe the trackers used in this Master Thesis: Template Matching (TM)[18][19], Mean-Shift (MS)[20], Particle Filter-based Colour tracking (PFC)[21], Lucas-Kanade tracking (LK)[22], Incremental learning for robust Visual Tracking (IVT) [23], Tracking Learning Detection tracking (TLD)[24], Corrected Background Weighted His-

---

<sup>1</sup> Some algorithm descriptions have been extracted (or edited) from [17].

togram tracker (CBWH) [25] and Scale and Orientation Adaptive Mean-Shift Tracking (SOAMST) [26]. The first four tracking algorithms have been chosen because they are classical and general tracking systems. The last four have been chosen because they are modern trackers with contrasted and remarkable results. Below is a summary describing each of them.

### 2.3.1 Template Matching (TM)

This single-target tracking algorithm[18][19] represents the target model by the subimage corresponding to the given rectangular region of interest to be tracked. The target model (template) is then searched over the current video frame by applying a convolution process in which the target model is the convolution mask. The candidate position is the location within the current frame with the largest convolution value. The convolution process can be replaced by other types of sum-comparing metrics, such as the sum of absolute differences (SAD), sum of square differences (SSD) and cross-correlation. Due to its inherent simplicity, this algorithm can be directly implemented in hardware or by taking advantage of vector machine code instruction sets (MMX, SSE, ...), hence making it suitable for real time processing. Its main drawback is its only invariance to translation changes of the target model, which can be the case for targets moving relatively slowly between consecutive frames. Notwithstanding, due to its extremely high computational efficiency, several templates can be generated by applying small rotations and scale changes to the original target model. The algorithm can then be applied to the different templates, finding out the one that yields the best matching with the current frame. Similarly, in order to cope with occlusions, the original target model can be partitioned into several templates that can then be independently matched with the current frame.

This tracking algorithm presents very good results in low complexity videos, specially where there are not appearance changes in the tracked target. If there are scale changes, rotations, similar objects or strong illumination changes, this algorithm fails.

### 2.3.2 Mean-Shift (MS)

This single-target tracking algorithm[20] represents the target model by the colour histogram of all pixels belonging to the given elliptical region of interest to be tracked. That histogram is computed in such a way that pixels close to the target center have a larger weight than those away from it according to the Epanechnikov kernel function. This weighting is done in order to lower the influence of pixels close to the boundaries of the region of interest, which are assumed to be less confident than those

close to the centre. The candidate position within the current video frame is the one that maximizes the Bhattacharyya distance between its associated colour histogram, which is computed in the same manner as the histogram of the target model, and the latter. That candidate position is found by iterating from the previously known target position until convergence by applying the mean-shift procedure to an image of weights. The larger the weight corresponding to a certain image position the larger the similarity between the colour histograms associated with both that position and the previously known target position. The algorithm can adapt to scale changes by slightly modifying the width of the Epanechnikov kernel function, thus slightly changing the area of the effective image region over which all histograms are computed. Three widths are considered: the previous width without changes and after both increasing it and decreasing it by 10%. The width that yields the maximum Bhattacharyya distance for the final candidate position is the one that denotes the change of scale. A simple variation of the algorithm described above, referred to as background-weighted histogram (BWH), aims at reducing the interference of background pixels in the tracking process by taking into account the colour histogram of the background surrounding the target model in order to modulate the colour histograms associated with the target model and the candidate positions. In particular, when a bin from the background histogram has a significant value, the corresponding bins for the target model and the candidate positions are given a low weight. The background histogram is computed in a region three times bigger than the area of the target model.

This algorithm works well in situations where color is a distinguishing feature of the tracked object. In sequences where there are objects with similar appearance, this algorithm usually loses the tracked object.

### 2.3.3 Particle Filter-based Colour tracking (PFC)

Similarly to the colour-based mean shift tracker summarized above, this single-target tracking algorithm[21] represents the target model by the colour histogram of all pixels belonging to the given elliptical region of interest to be tracked. That histogram is also computed in such a way that pixels close to the target's centre have a larger weight than those away from it according to the Epanechnikov kernel function. However, differently to the mean shift tracker, the candidate position of the target model in the current video frame is found as a weighted average of alternative candidate positions, each referred to as a particle. Every particle is represented by the position, size and the corresponding first derivatives of a 2D ellipse. The weight associated with each particle is computed according to the Bhattacharyya distance between the

colour histograms of both the target model and the ellipse corresponding to that particle, such that the larger the distance, the larger the weight. Every particle iteratively evolves at every time step by changing its position and size according to its corresponding first derivatives plus a random offset following a zero-mean Gaussian distribution. The derivatives are also changed by applying a random offset. Initially, all particles can be randomly distributed over the video frame in order to cover regions where the target is expected to appear or where an object detection algorithm determines. The iterative algorithm stops when the candidate position converges.

This tracker performs better in complex sequences due to the way in which the various possible positions are considered. Therefore, this is one of the most used trackers from the state of the art.

#### 2.3.4 Lucas-Kanade tracking (LK)

This single-target tracking algorithm[22] can be considered to be a generalization of the above template matching algorithm that allows for small affine transformations (translation, rotation, scaling, shear mapping, etc.) of the target model. In particular, the target model is represented by the subimage corresponding to the given rectangular region of interest to be tracked. The target model (template) is then searched over the current video frame by finding the parameters of the affine transformation that best aligns the transformed image with the target model. That search is cast as a minimization problem that is iteratively solved by applying gradient descent, starting with an initial estimation of the sought parameters. Since the variation between consecutive video frames is usually small, this initial estimation can simply be the values of the parameters corresponding to the target model in the previous frame or zeroes for the first frame.

In sequences where there are slow appearance changes, this tracker works better than the others. Despite this, this tracker presents difficulties in many of the problems that a tracker can face: complex movements, illumination changes, occlusions, etc.

#### 2.3.5 Incremental learning for robust Visual Tracking (IVT)

This single-tracking algorithm[23] incrementally learns a low dimensional eigenbasis representation, adapting online to changes in the appearance of the target. The model update, based on incremental algorithms for principal component analysis, includes two features: a method for updating the sample mean, and a forgetting factor to ensure less modelling power is expended fitting older observations. The incremental eigenbasis learning approach exploits the local linearity of appearance manifold for

matching targets in consecutive frames. Whereas most algorithms operate on the premise that the object appearance or ambient environment lighting conditions do not change as time progresses, this algorithm adapts the model representation to reflect appearance variation of the target, thereby facilitating the tracking task. In contrast to the existing incremental subspace methods, the eigenbasis method updates the mean and eigenbasis, and thereby learns to model the appearance of the target being tracked. This tracker occasionally drifts from the target object. With the help of particle filters, the tracker often recovers from drifts in the next few frames when a new set of samples is drawn. For specific applications, better mechanisms to handle drifts could enhance robustness of the algorithm.

This tracking algorithm works well when the model learned is updated correctly. In cases where the appearance change is not linear (i.e. very fast), the target can get lost. Once the object is lost, the particle filter can not recover the object due to the previous appearance changes.

### **2.3.6 Tracking Learning Detection tracking (TLD)**

This single-target long term tracking algorithm[24] can be seen as a combination of tracking and detection. The tracking component estimates the object motion between consecutive frames under the assumption that the frame to frame motion is limited and the object is visible. The tracker is likely to fail and never recover if the object moves out of the camera view. Detector treats every frame as independent and performs full scanning of the image to localize all appearances that have been observed and learned in the past. As any other detector, the detector makes two types of errors: false positives and false negatives. Learning observes performance of both, tracker and detector, estimates detector's errors and generates training examples to avoid these errors in the future. The learning component assumes that both the tracker and the detector can fail. The key idea of the learning is that the detector errors can be identified and corrected.

This tracker works well when the object tracking is done correctly in the first frames: the longer the object is correctly followed in the first frames, the better the learning. In the cases where tracking drift occurs during the first frames, the learned model is incorrect and the performance of the tracker is bad.

### **2.3.7 Corrected Background-Weighted Histogram tracker (CBWH)**

This single-target tracking algorithm[25] is a variation of the original colour-based mean-shift technique [20] that modifies the stage that reduces the interference of

background pixels, originally referred to as background-weighted histogram (BWH). In particular, the proposed algorithm, referred to as corrected background-weighted histogram (CBWH) only transforms the histogram of the target model, but not the histograms of the candidate positions, thus decreasing the probability of target model features that are prominent in the background. Experimental results show that CBWH can reduce the number of mean-shift iterations, as well as improve the tracking accuracy. One of its main advantages is that it reduces the sensitivity of mean-shift tracking to the target initialization. Therefore, CBWH can robustly track the target even if it is not initialized precisely.

In sequences in which the algorithm CBWH is able to correctly differentiate the background and the object, the algorithm will present good results. In the cases where there are similar objects near the tracked object, the algorithm has difficulties.

### 2.3.8 Scale and Orientation Adaptive Mean-Shift Tracking (SOAMST)

This single-target tracking algorithm[26] is a variation of the original colour-based mean-shift technique [20] that is able to update the scale and orientation of the target model during the tracking process. The original mean-shift tracker only supports discrete changes in the scale of the target model. In the proposed variation, the image of weights generated by the original mean-shift tracker, in which a pixel has a large weight if the colour histogram associated with that candidate position is similar to the histogram of the target model, is utilized to estimate the area and orientation of the target. In particular, the zero-th-order moment of the image of weights is utilized to estimate the area of the target model, and hence its scale, whereas the width, height and orientation changes of the target are estimated using the area estimated before, as well as the second-order centre moment of the image of weights.

This algorithm has the advantages of the Mean Shift approach, adding support scale and orientation changes. Despite this, the added features can cause some Mean Shift tracking errors that were not produced in the original algorithm.

## 2.4 Fusion

Multiply and as much independent as possible sources of information are commonly used in signal processing to improve the result of an algorithm. Using multiple independent features usually improves the performance and the robustness of a video tracker. The fusion strategies (for video object tracking) can be classified in two different architectures[27][4], parallel and sequential, and in two main levels[4], tracker-level fusion and measurement level fusion.

### 2.4.1 Fusion architectures

In the parallel architecture, each tracker is executed independently and the result of the fusion is the most confident tracker or the best combination of trackers. In the sequential or cascade architecture, trackers are evaluated sequentially: if the first tracker returns a high confidence<sup>2</sup>, its result is chosen as the fusion result; otherwise, the next tracker is evaluated and the process applied is repeated.

### 2.4.2 Fusion levels

Fusion at tracker level models single-feature tracking algorithms as black boxes. The video tracking fusion problem is redefined by modeling the interaction between outputs of each tracker, which can run in parallel or in cascade (sequentially). Fusion can use classical combination techniques (average, maximum, minimum, median, ...)[28], combine the resulting Probability Density Function (PDF) of each algorithm[29], consider variable weights for the algorithms[29][30], use a probabilistic approach[31][32], add a later prediction stage[33], combine the resulting bounding box of each tracker[34][32] or combine results at pixel (segmentation) level[16]. For the fusion of multiple features at measurement level, the measurements are combined internally by the tracking algorithm. Measurement level fusion can take place with a variety of mechanisms, such as using Bayesian methods[35][36], particle filtering[37][23], estimating mutual information[38] or calculating correlation[39].

### 2.4.3 Combination techniques

There are many classical combination techniques that can be applied to object tracking, as those described in [28], where a detailed study of several possible combinations is presented. Although this reference is not focused on tracking objects, many of the combinations presented can be applied to it. The main classic combination techniques presented are majority vote, weighted majority vote, naive bayes combination, multinomial methods and other approximations such as those mentioned previously (mean, median, ...). For each of these techniques a detailed analysis is presented.

Using variable weights for the algorithms is a common technique used in object tracking. In [29], the trackers are considered as black boxes and the combination uses only the trackers output, which may be modified before their propagation to the next time step (feedback). A probabilistic framework is proposed for combining multiple synchronous trackers, where each separate tracker outputs a PDF of the tracked state. In the approach presented in [30], the ensemble of weak classifiers is combined into

---

<sup>2</sup>This confidence measure must be given by the tracker



a strong classifier using AdaBoost. The strong classifier is then used to label pixels in the next frame, giving a confidence map. Pixels can belong to the object or to the background. The peak of the map, calculated via Mean Shift, is the place where the object is supposed to be. In the update state, the algorithm keeps the best weak classifiers (updating their weights) and adds new classifiers.

A probabilistic approach is used in [31] to infer the most likely object position and the accuracy of each tracker. A testing sample is chosen to be the new object position if it has maximum probability and if it belongs to a positive sample. In each iteration, the target appearance model is updated with an on-line learning method. This method tries to avoid the tracking drift problem (a gradual adaption of the tracker to the background instead of to the target). In the case of [32], a crowdsourcing tracking method is proposed to infer simultaneously the ground truth bounding box of the tracked object and the confidence for each tracker. Then, a particle filter technique is used to approximate the a-posteriori density.

A classification technique is used in [33] to detect tracking errors. Only the tracker results classified as correct are fused. An additional prediction stage is added to the system with a Kalman filter.

As the previous works presented, [34] uses only information from the output of each considered tracker. In this case, the information used is obtained from the bounding boxes generated by each one of the trackers. A Gaussian Model is formulated for the center and for the vertical and horizontal information of the bounding box. The resulting bounding box is obtained selecting the tracker with the maximum confidence after combining the information of all the trackers.

Another perspective to address the tracking fusion problem is to consider it at pixel level, as reported in [16]. The main idea is to transform the different trackers outputs as motion inliers, bounding boxes or specific target image features to a unique representation, on which the fusion step will be applied. This combination is done at pixel level, generating a final segmentation obtained by combining information from all the trackers. The tracked object position is estimated using this segmentation.

For the measurement level fusions, the approach presented in [37] is based on the use of a particle filter. This multi-feature fusion model combines color and edge orientation by a stochastic fusion scheme. Observation models are statistical models describing occurrences of features. The color feature is based on a HSV color histogram, used to compute the likelihood of color. The edge orientation feature is obtained convolving the grayscale intensity images with Sobel masks. Another use of Particle Filters is presented in [23], proposing a method that incrementally learns a low-dimensional subspace representation of the appearance of the object: the learning

algorithm draws particles in a motion parameter space and predicts the best location of the tracked object using information from the appearance model.

Another example of measurement level fusion is the work presented in [35]. This work focuses on achieving robustness against appearance and motion changes. To achieve it, the observation model is decomposed into multiple simpler observation models generated by Sparse Principal Component Analysis (SPCA). Each one of the generated models covers a specific appearance of the object. The motion model is also represented with a combination of multiple basic motion models. In [36], the work presented in [35] is continued and improved. This improvement is based on searching for the trackers which work correctly in each frame. This combination technique is called visual tracker sampler. This method obtains multiple samples of the trackers using the Markov Chain Monte Carlo method. After each Markov Chain is modeled, they run in parallel and produce samples of the states to estimate each decomposed a-posteriori probability. When the chains are in the interacting mode, they communicate with the others and leap to better states of the target. During the sampling process, the number of Markov Chains changes by either increasing or decreasing the number of considered trackers.

[38] and [39] are not object tracking fusion works, but their ideas may be applied to this research field. The main idea presented for combining independent data sources is to use the correlation obtained between each pair of information sources, so that the final decision is the combination of information sources that have higher correlation. This method eliminates the need for supervised annotations or feedback. In [38], the used agreement measure is the Kendall  $\tau$ , while in [39] the similarity measure proposed to use is the classical correlation metric between two signals, taking advantage of its simplicity and computational efficiency.

## 2.5 Conclusions

There are multiple types of video object tracking systems: multicamera, multi-target, monocamera, multisensor, etc. The presented work is only centered in single-tracking monocamera algorithms. These are the simplest algorithms on which the other more complex cases can be designed.

Each tracking algorithm has its advantages and disadvantages, depending on the environment in which it runs. In chapter 4, the eight individual algorithms described in this chapter will be evaluated. The chapter will present the results and verify constraints and performance of each of them.

Due to the specificity of the algorithms, the fusion of trackers is an interesting

aspect to consider, in order to obtain the generality absent in individual algorithms. In this chapter, we have presented some of the state of the art fusion methods. Chapter 5 describes some fusion methods implementations and their results.



## Chapter 3

# Evaluation Framework

### 3.1 Introduction

The evaluation of object tracking algorithms is necessary to validate its correctness and robustness. This chapter presents the content set and the metrics considered for the development of this Master Thesis. The content set used was created by the VPULab trying to independently address the different problems that a tracker can face. After that, a set of metrics, obtained from the state of the art, is presented. Additionally, appendix 3.3.2 presents a study of the correlation between the different metrics considered in this chapter.

### 3.2 Single Object Video Tracking dataset - SOVTds<sup>1</sup>

For video object tracking, the SOVTds was created by the VPULab focused on the main problems that affect video object tracking in surveillance videos. Moreover, a description of publicly available datasets is also provided in the appendix A.

The selection of the test scenarios is one of the most important steps when developing an evaluation protocol. Each issue has to be represented in the dataset for achieving a correct understanding of the capabilities of the tracking algorithm. Moreover, different levels of complexity have to be covered in the test data. Hence, this dataset was designed[17]with four complexity levels including both real and synthetic sequences. The addressed problems and the modeled situations are described as follows.

---

<sup>1</sup>This section has been edited from [17].

### 3.2.0.1 Selected tracking problems

Several problems have to be taken into account that corresponds to real-world situations. In the SOVT dataset, the following tracking-related problems have been modeled:

1. Complex (fast) motion: the target changes its trajectory unexpectedly or increases its speed abruptly; the tracker might lose the target if it exceeds the search area.
2. Gradual (and global) illumination changes: in long sequences, the illumination might change due to weather conditions, time passing, etc. In this case, the target model might become outdated making harder the tracking task.
3. Abrupt (and local) illumination changes: as the target moves, it can enter in areas with different illumination. Hence, the tracker might be confused and lose the target.
4. Noise: it appears as random variations over the values of the image pixels and can significantly degrade the quality of the extracted features for the target model.
5. Occlusion: it is defined when an object moves between the camera and the target. It can be partial or total if, respectively, a region or the whole target is not visible.

### 3.2.0.2 Complexity factors

In table 3.1, the criteria for defining the complexity factors of the test sequences are described.

### 3.2.0.3 Modeled situations

As a tracker can operate in different conditions in which the same problem appears, the sequences are organized into four situations ranging from completely controlled (e.g., synthetic sequences) to uncontrolled (e.g., real-world sequences). Moreover, the complexity of the tracking problems is estimated for each sequence of the situations. The complexity-level sequences sets are:

1. Synthetic sequences set (S1): it is composed of synthetic sequences that provide controlled testing conditions allowing to isolate each problem. They consist on a moving ellipse in a black background that can contain squares of the same or

Problem	Criteria (factors)
Complex Movement	The target changes its speed (pixels/frame) abruptly in consecutive frames
Gradual Illumination	The average intensity of an area changes gradually with time until a maximum intensity difference is reached
Abrupt Illumination	The average intensity of an area changes abruptly with respect to its surroundings (maximum intensity difference)
Noise	It includes natural (snow) or white Gaussian noise which is manually added with varying deviation value
Occlusion	Objects in the scene occlude a percentage of the target
Scale Changes	The target changes its size with a maximum relative change regarding its original size
Similar Objects	An object with similar color to the target appears in the neighborhood of the target

Table 3.1: Complexity factors for the video tracking dataset

different color (acting as, respectively, similar or occlude objects). The created sequences to model all the selected problems have five degrees of complexity for each one. In total, 35 sequences were generated with around 3500 frames. Sample frames are shown in the first row of figure 3.1.

2. Laboratory sequences set (S2): it provides a natural extension of the S1 situation by representing real test data in a laboratory setup under controlled conditions. An object with a simple color pattern was used for generating such data. These sequences have been recorded to model all the selected problems with three complexity levels for each one. For some problems (complex movement, occlusion, scale changes and similar object problems), the sequences were recorded using the test object, whereas for the other ones (noise, gradual and abrupt illumination changes), a single sequence was recorded without any problems and then they were artificially included. In total, 21 sequences were generated with around 6500 frames. Sample frames are shown in the second row of Figure 3.1.
3. Simple real sequences set (S3): it includes data from previously existing datasets that have been captured in non-controlled conditions. Clips have been extracted from the original sequences that contain isolated tracking problems. As each target has different characteristics [4], the sequences have been grouped into three target-dependent categories: cars (from MIT Traffic [40] and Karlsruhe [41] datasets), faces (from TRECVID2009 [42], CLEMSON[43] and VISOR [44]

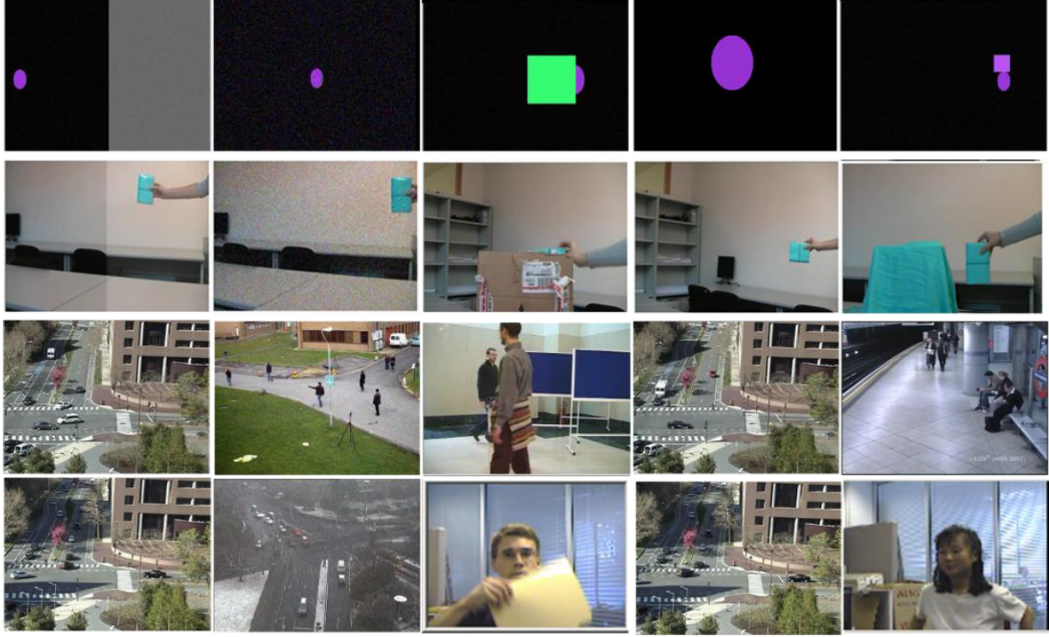


Figure 3.1: Sample frames for the situations of the proposed dataset (from top row to bottom row): Synthetic set (S1), Laboratory set (S2), Simple real set (S3) and Complex real set (S4). In addition, samples of some tracking-related problems are also presented for each column (from left to right): abrupt illumination change, noise, occlusion, scale change and (color-based) similar objects.

datasets) and people (from TRECVID2009 [42], i-Lids [45], PETS2009 [46], PETS2000 [46] and CAVIAR [47] datasets). For each target type and problem, three sequences with varying complexity level were composed making a total of 53 sequences with around 8500 frames. Sample frames are shown in the third row of Figure 3.1.

4. Complex real sequences set (S4): the last set contains the most complex sequences, which are clips from other datasets that include several problems. Once the algorithms are tested for each problem individually, it is a good idea to check the performance in more realistic (and complex) situations. Similarly to the previous situation, we also distinguish three problems which have been estimated and classified according the defined criteria. All these sequences were extracted from the MIT Traffic [40] (for cars), CLEMSON [43] (for faces) and PETS2009 [46] (for people) datasets. In total, 15 sequences were selected with around 4500 frames. Sample frames are shown in the fourth row of Figure 3.1.



### 3.3 Selection of evaluation metric

#### 3.3.1 Metrics<sup>2</sup>

In this section, we describe the metrics that have been used in this Master Thesis for performance evaluation of single-object video tracking. .

In subsection 3.3.2, a metric correlation study is presented showing redundancy between many of the metrics presented. This study, based on the results of chapter 4, justifies the decision to use a single metric, namely SFDA (see section 3.3.1.1), to compare the different tracking approaches.

We will first begin by introducing the common base metrics for describing complex metrics:

- $TP_P$ : True positive, a target pixel appears both in the ground-truth annotation and the algorithm result (per frame).
- $TN_P$ : True negative, a target pixel that appears neither in the ground-truth annotation nor the algorithm result (per frame).
- $FP_P$ : False positive, a target pixel that appears in the algorithm result, but not in the ground-truth annotation (per frame).
- $FN_P$ : False negative, a target pixel that appears in the ground-truth annotation but not in the algorithm result (per frame).

##### 3.3.1.1 Sequence Frame Detection Accuracy (SFDA)

The Sequence Frame Detection Accuracy (SFDA)[48] measure calculates in each frame the spatial overlap between the estimated target location and the ground-truth annotation. This mapping is optimized on a frame-by-frame basis. It contains information regarding the number of objects detected, missed detections, false positives and spatial overlap, providing a ratio of the spatial intersection and union between two object locations. The total sum of data from the Frame Detection Accuracy (FDA) is then normalized to the number of frames including ground-truth targets. Therefore, SFDA can be seen as the average of the FDA over all the relevant frames in the sequence. SFDA ranges from 0 to 1; the higher the value, the better.

$$\mathbf{SFDA} = \frac{\sum_{t=1}^{t=Nframes} FDA(t)}{\sum_{t=1}^{t=Nframes} \exists(N_G^{(t)} OR N_D^{(t)})} \quad (3.1)$$

---

<sup>2</sup>This subsection is based and extends part of the work presented in [1]

$$FDA(t) = \frac{Overlapratio}{N_G^{(t)} + N_D^{(t)}} \quad (3.2)$$

$$Overlapratio = \sum_{i=1}^{N_{mapped}^{(t)}} \frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|} \quad (3.3)$$

Where:

$G_i^{(t)}$  denotes the i-th ground-truth object in frame t.

$D_i^{(t)}$  denotes the i-th detected object in frame t.

$N_G^{(t)}$  and  $N_D^{(t)}$  denote the number of ground-truth objects and the number of detected objects in frame t, respectively.

$N_{frames}$  is the number of frames in the sequence.

$N_{mapped}^{(t)}$  is the number of mapped ground truth and detected object pairs in frame t (frame level mapping).

### 3.3.1.2 Average Tracking Accuracy (ATA)

The Average Tracking Accuracy (ATA)[48] is a spatiotemporal measure that penalizes fragmentations in both the temporal and spatial dimensions while accounting for the number of objects detected and tracked, missed objects and false positives. A one-to-one mapping between the ground truth and the system output objects was established by computing the measure over all of the ground truth and detected object combinations and using an optimization strategy to maximize the overall score for the sequence. The Sequence Track Detection Accuracy (STDA) is calculated for ATA. STDA corresponds to the SFDA when there is a matching between ground-truth annotation and the estimated target location. ATA ranges from 0 to 1; the higher the value, the better.

$$\mathbf{ATA} = \frac{STDA}{\frac{N_G + N_D}{2}} \quad (3.4)$$

$$STDA = \sum_{i=1}^{N_{mapped}} \frac{\sum_{t=1}^{N_{frames}} \frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|}}{N_{(G_i \cup D_i \neq 0)}} \quad (3.5)$$

Where:

$G_i^{(t)}$  denotes the i-th ground-truth object in frame t.

$D_i^{(t)}$  denotes the i-th detected object in frame t.

$N_G^{(t)}$  and  $N_D^{(t)}$  denote the number of ground-truth objects and the number of detected objects in frame t, respectively.

$N_{frames}$  is the number of frames in the sequence.

$N_{mapped}$  is the number of mapped ground truth and detected object pairs at the sequence level.

### 3.3.1.3 Average Tracking Error (ATE)

The Average Tracking Error (ATE)[49] can be seen as a false positive rate (whereas ATA represents the true positive rate). It provides a ROC-like curve which allows comparing and evaluating the tracker's performance. ATE ranges from 0 to 1; the lower the value, the better.

$$\mathbf{ATE} = \frac{1}{N_{frames}} \sum_{i=1}^{N_{mapped}} \frac{|D^{(t)} \setminus G^{(t)}|}{|D^{(t)}|} \quad (3.6)$$

Where:

$N_{frames}$  is the number of frames in the sequence.

$N_{mapped}$  is the number of mapped ground truth and detected object pairs at the sequence level.

$|D_t \setminus G_t|$  is the relative complement, that is, the set of elements in  $D_t$ , but not in  $G_t$ .

### 3.3.1.4 Overlap

This measure[50] determines the amount of overlap between the ground-truth annotations and estimated target locations. It is computed in every frame where the target exists. Overlap ranges from 0 to 1; the higher the value, the better.

$$\mathbf{Overlap} = \frac{TP_p}{TP_p + FP_p + FN_p} \quad (3.7)$$

### 3.3.1.5 Area Under the lost track ratio Curve (AUC)

For the AUC[50] metric, a target is said to be lost when the spatial overlap between the ground-truth and the estimated target is smaller than a threshold. Afterward, the lost-track ratio ( $\lambda$ ) is calculated based on the overlap of the sequence. Because the appropriate value of the threshold  $\tau$  is different for different tracking applications, it was considered the variation of  $\tau$  for a full range of values: from  $\tau = 0$  to  $\tau = 1$  with an increment of 0.01. We refer to these values of the lost-track ratio as  $\lambda(\tau)$ . AUC ranges from 0 to 1; the lower the value, the better.

$$\mathbf{AUC} = \Delta\tau \sum_{\tau=0}^1 \lambda(\tau) \quad (3.8)$$

$$\lambda = \frac{N_l}{N} \quad (3.9)$$

Where:

$N_l$  is the number of frames with a lost track. A track in a frame  $t$  is considered to be lost when the amount of overlap (see subsection 3.3.1.4) between the estimated track and the ground truth is smaller than a certain value  $\tau$ .

$N$  is the total number of frames of the estimated target trajectory.

### 3.3.1.6 Closeness of Track (CT)

This metric[51] aims to calculate the average closeness between a pair of ground-truth and system results tracks. The closeness of the whole sequence (CTM) can be averaged by weighting the CT of all pairs. The weighted standard deviation of track closeness (CTD) can be also obtained for the whole sequence. CTM ranges from 0 to 1; the higher the value, the better.

$$\mathbf{CTM} = \frac{\sum_{t=1}^M CT_t}{\sum_{t=1}^M \text{length}(CT_t)} \quad (3.10)$$

$$\mathbf{CTD} = \frac{\sum_{t=1}^M \text{length}(CT_t) \times \text{std}(CT_t)}{\sum_{t=1}^M \text{length}(CT_t)} \quad (3.11)$$

$$CT(G_i, D_i) = \{A(G_i^{(1)}, D_i^{(1)}), \dots, A(G_i^{(t)}, D_i^{(t)})\} \quad (3.12)$$

Where:

$A$  represents the spatial overlap for ground truth and system tracks.

$G_i^{(t)}$  denotes the  $i$ -th ground-truth object in frame  $t$ .

$D_i^{(t)}$  denotes the  $i$ -th detected object in frame  $t$ .

### 3.3.1.7 Track Completeness (TC)

This metric[51] is defined as the time span that there is overlap between the system and ground-truth tracks and divided by the duration of the ground truth track. TC ranges from 0 to 1; the higher the value, the better.

$$\mathbf{TC} = \frac{\sum_{t=1}^{N_D^{(t)}} O(G^{(t)}, D^{(t)})}{N_G^{(t)}} \quad (3.13)$$

Where:

$O(G^{(t)}, D^{(t)})$  is a binary variable with value 1 if a pair is overlapped more than a

threshold value (we used  $th = 0.2$ ) and 0 otherwise.

$N_D^{(t)}$  is the duration of the Detection track.

$N_G^{(t)}$  is the duration of the Ground Truth track.

### 3.3.1.8 Combined Tracking Performance Score (CoTPS)

This metric[52] is based on a previous defined metric, AUC (see section 3.3.1.5), and combines information of tracking accuracy and tracking failure in a single score to facilitate performance ranking. CoTPS ranges from 0 to 1; the lower the value, the better.

$$\mathbf{CoTPS} = \beta \cdot AUC + (1 - \beta) \cdot \lambda_0 \quad (3.14)$$

$$\lambda_0 = \frac{N^0}{N} \quad (3.15)$$

$$\beta = \frac{\hat{N}}{N} \quad (3.16)$$

Where:

$N^0$  is the number of frames in which there is no overlap between the ground truth and the detection.

$N$  is the number of frames in which the object is tracked.

$\hat{N}$  is the number of frames in which the overlap between the ground truth and the detection is higher than a threshold.

$AUC$  is the Area Under the lost track ratio Curve.

### 3.3.2 Metrics correlation study

This section presents a correlation study among all the metrics described in section 3.3.1. For this correlation study the CTD metric has not been considered since it differs from the other metrics (CTD represents a weighted standard deviation). The remaining metrics values range between 0 and 1: the higher the value, the better the tracking result.

For the correlation study, all the scores obtained after processing each video of the dataset are used to generate a vector of 976 values (122 videos x 8 trackers) for each one of the metrics. Table 3.2 presents the correlation coefficient between each pair of metrics.

	SFDA	ATA	ATEinv	AUCinv	Overlap	CTM	TC	CoTPSinv
SFDA	1,00	1,00	0,64	0,96	0,96	1,00	0,89	0,88
ATA	1,00	1,00	0,64	0,96	0,96	1,00	0,89	0,88
ATEinv	0,64	0,64	1,00	0,60	0,60	0,64	0,61	0,42
AUCinv	0,96	0,96	0,60	1,00	1,00	0,96	0,89	0,92
Overlap	0,96	0,96	0,60	1,00	1,00	0,96	0,89	0,92
CTM	1,00	1,00	0,64	0,96	0,96	1,00	0,89	0,88
TC	0,89	0,89	0,61	0,89	0,89	0,89	1,00	0,78
CoTPSinv	0,88	0,88	0,42	0,92	0,92	0,88	0,78	1,00

Table 3.2: Correlation between metrics

As can be seen in table 3.2, all metrics show a similar correlation (around 0.9) in most cases. There is only one exception with the ATEinv metric. As explained in its definition (see section 3.3.1.3), the ATE metric can be seen as a false positive rate. This means that this measure, unlike the rest, does not penalize the existence of false negatives, as it only considers the number of false positives.

After obtaining these results, a single metric has been chosen to sum results: SFDA. Only this metric will be used to draw conclusions in most chapters, thus avoiding redundant information. This metric has been chosen for two main reasons: its correlation with respect to other metrics (excluding, as discussed above, the metric ATE) is one of the the highest. SFDA also considers and penalizes both false positives and false negatives. Probably any of the metrics that have higher correlation with the SFDA metric(ATA, AUC,...), would have obtained similar (or even identical) conclusions.

### 3.4 Conclusions

Although there are several datasets (see appendix A), we have selected the SOVTds because it was created trying to independently address the different problems that a tracker can face: complex movement, local and global illumination, noise, oclusions, scale changes and similar objects. Each of these problems are addressed by three different degrees of difficulty, corresponding to the first three categories of the dataset.

About the metrics, there are several possibilities in the state of the art. Nine different metrics from the state of the art have been presented and defined. As can be seen in subsection 3.3.2, most of these metrics are highly correlated so one single metric (we have selected the SFDA metric) can be chosen to extract general conclusions without losing precision.

## Chapter 4

# Individual trackers evaluation

### 4.1 Introduction

In this chapter we present the results of the evaluation for each individual tracker and some discussions of their performance. After that, comparative results are presented to analyze which trackers perform worse or better in certain environments. These results and conclusions are based on the SFDA metric (see subsection 3.3.2). The complete results for all metrics can be found in appendix B and C.

### 4.2 Individual Tracking Algorithms

#### 4.2.1 Template matching (TM)

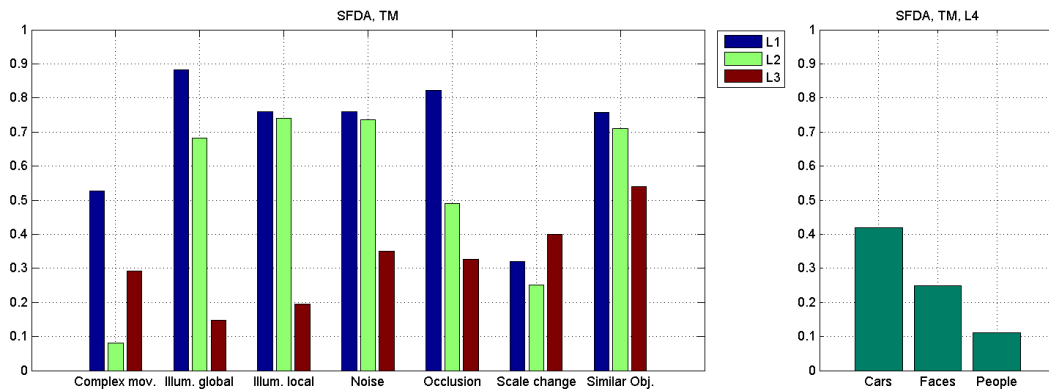


Figure 4.1: SFDA result for TM tracker

Due to the relative simplicity of the TM algorithm, its best scores are obtained in the sequences belonging to L1 and L2, which correspond to the simplest tracking sequences. This algorithm does not consider scale changes, what is reflected in the decrease in performance for sequences of this subcategory (*scale change*) in the three first sets of sequences. For the L1 sequences, the TM algorithm gets the highest score in *illumination global* (along with IVT), *illumination local*, *noise*, *occlusions* and *similar object* subcategories, and the second best score in *complex movement* categories. In most L1 and L2 sequences, the object suffers very slight changes in appearance (L1 synthetic and L2 rigid object) and, in general, the tracking is performed correctly using the appearance (template) of the first frame.

#### 4.2.2 Mean-Shift (MS)

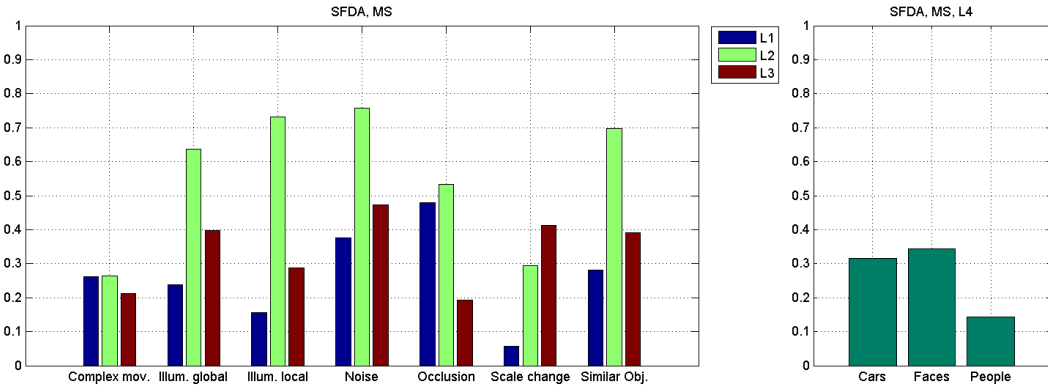


Figure 4.2: SFDA result for MS tracker

The MS algorithm does not stand out in any of the categories of the sequences L1, L3 and L4. Its results are generally improved by the SOAMST and the CBWH trackers, since these latter are designed based on the MS algorithm with certain improvements. For the L2 sequences, the MS tracker obtains notable results in *illumination global*, *illumination local*, *noise*, *occlusions* and *similar objects* subcategories. As this algorithm represents the target model by the color histogram, in L2 sequences the differentiating color (turquoise) of the tracked object allows discriminating it with relative ease in the sequences.



### 4.2.3 Particle Filter-based Colour tracking (PFC)

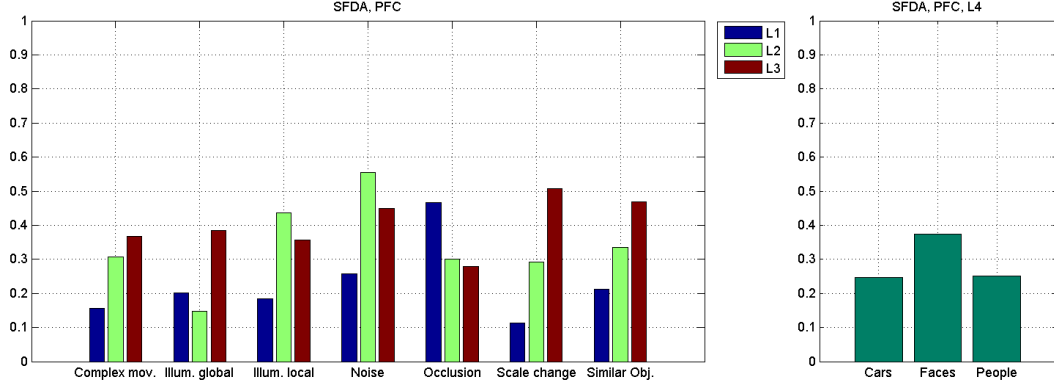


Figure 4.3: SFDA result for PFC tracker

In the case of the PFC algorithm, for the first sequences set (L1) the presented results are the worst for most of the categories. The PFC algorithm does not work properly in synthetic sequences, because uniform regions cause malfunctions in the particle filter. For categories L2 and L3, the algorithm does not get any remarkable result. L4 is the category in which this algorithm shows the best performance, achieving the best score in two of the three subcategories (*faces* and *people*). Thanks to the alternative candidate positions (particles) this algorithm performs better in environments with higher complexity, where others trackers fail.

### 4.2.4 Lucas-Kanade tracking (LK)

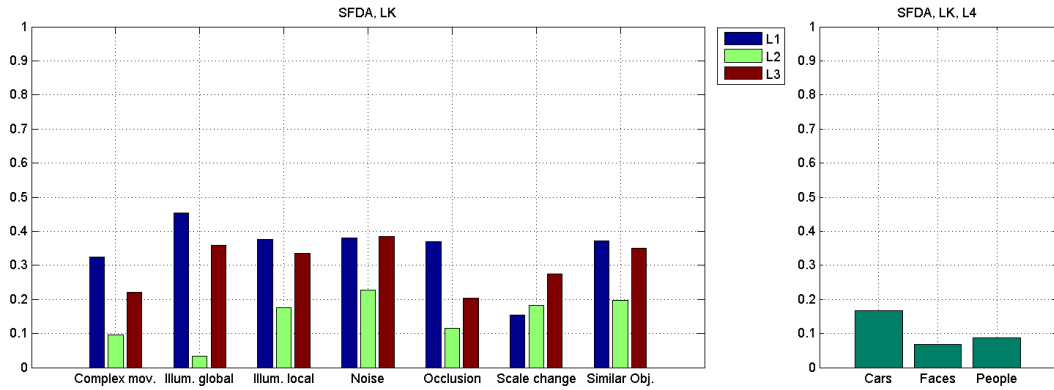


Figure 4.4: SFDA result for LK tracker

The results obtained by the LK algorithm for the L1 sequences are far from the best results obtained in each of the seven subcategories. In the case of the L2 sequences, results in each category are the worst of all, except in the *complex movement* and in the *illumination global* category. For L3 and L4 sequences, the LK algorithm presents average results compared with the other trackers. As was mentioned in subsection 2.3.4, this tracker presents difficulties in many of the problems that a tracker can face: complex movements, illumination changes, occlusions, etc. Besides, appearance changes that occur in most sequences are greater than the appearance changes that the algorithm is able to estimate.

#### 4.2.5 Incremental learning for robust visual tracking (IVT)

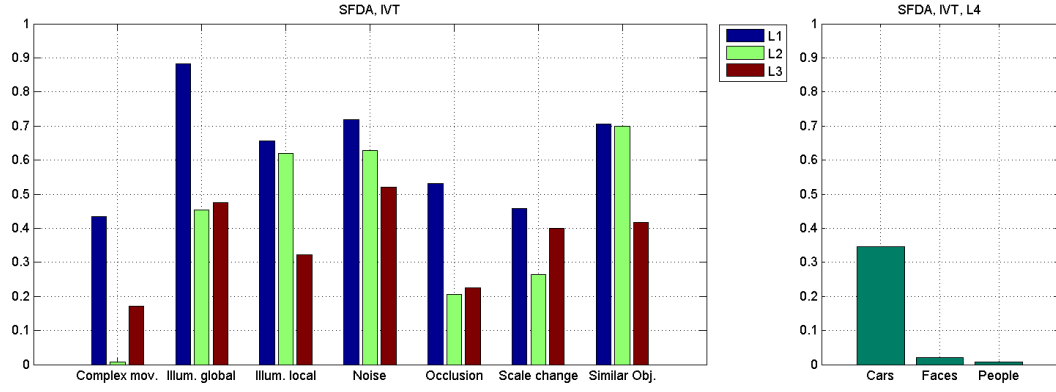


Figure 4.5: SFDA result for IVT tracker

IVT algorithm shows good results (top 3) for L1 sequences. For L2 sequences, this algorithm shows average results except for the subcategory of *complex movement*, where IVT stands out negatively probably because in that sequences the learning is done incorrectly due to the movements of the tracked object, getting the worst score of all. The same applies to the L4 category, where it obtains the worst results for *faces* and *people* sequences. In the case of the L3 category, the algorithm obtains average results except for *global illumination* and *noise* subcategories, obtaining the second best scores. The poor performance of this algorithm is because if the learning of the first frames suffers errors, the proper functioning of the rest of the sequence gets significantly complicated (tracking drift).

#### 4.2.6 Tracking learning detection tracking (TLD)

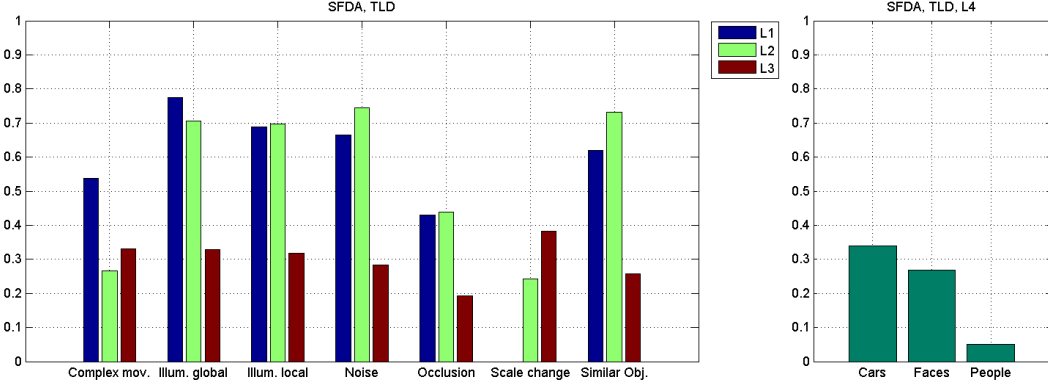


Figure 4.6: SFDA result for TLD tracker

The TLD tracker gets its best performance in the categories L1 and L2. It obtains the best score in L1 *complex movement*, L2 *illumination global* and L2 *similar object* subcategories. For most of the remaining subcategories (including L3 and L4), medium-high results are obtained. The failure of robustness obtained in L1 *scale change* subcategory should be noted. As the previous tracker, this tracker is based on the learning of the object model. In contrast to the previous tracker, the learning is accomplished properly and that is why good results have been achieved, especially in L1 and L2 sequences where the tracked object appearance changes are smaller and the object model learned is appropriate.

### 4.2.7 Corrected Background-Weighted Histogram tracker (CBWH)

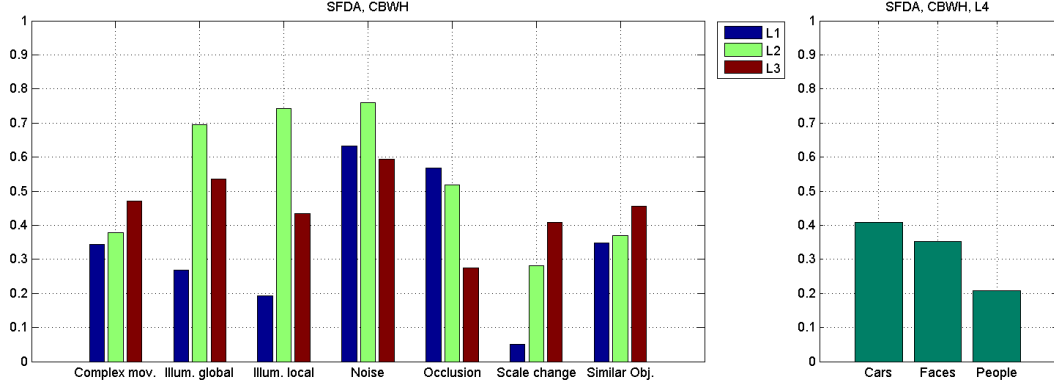


Figure 4.7: SFDA result for CBWH tracker

The CBWH algorithm presents average performance for L1 sequences. The best performance of this algorithm is obtained in the L3 category, where it gets the best score in four of the seven sub-categories: *complex movement*, *illumination global*, *illumination local* and *noise*. For L4 category, CBWH algorithm gets its scores in the top three of the three subcategories. Thanks to the capacity of this tracker to reduce the effect caused by the background in the tracking initialization, the object is tracked correctly in more complex sequences thanks to the object model is less distorted than the model created by other algorithms like template matching or mean shift, which do not consider this effect.

### 4.2.8 Scale and Orientation Adaptive Mean-Shift Tracking (SOAMST)

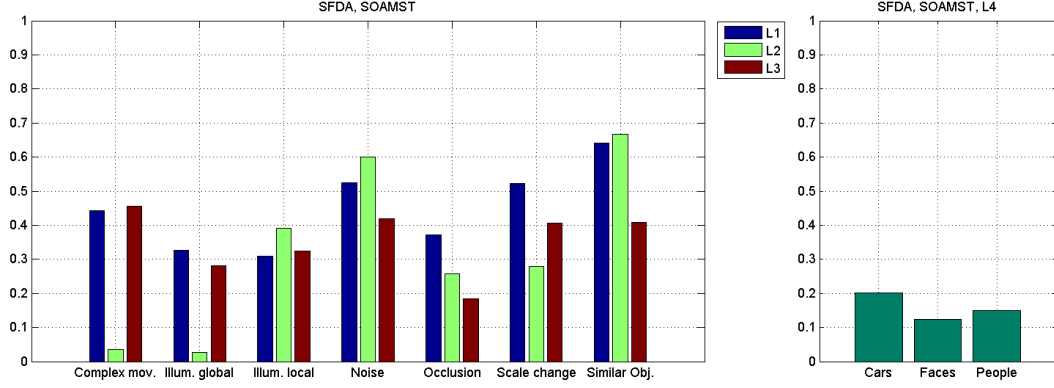


Figure 4.8: SFDA result for SOAMST tracker

For L1 sequences, the SOAMST algorithm has better performance than the others algorithms in the *scale change* category. This algorithm has been specially designed to withstand scale and orientation changes. Despite being the most modern algorithm, SOAMST algorithm does not stand out in any of the subcategories of L2, L3 and L4, even in the *scale change* subcategories. As this algorithm considers more complex situations, in some sequences it believes that there are changes in the sequence that does not happen, and therefore fails. This problem causes malfunction even for sequences for which this tracker has been specifically designed (*scale changes* of L2 and L3 sequences).

## 4.3 Comparative results

This subsection presents the comparative results of the algorithms. As in the previous sections, the metric used for the comparison is the SFDA.

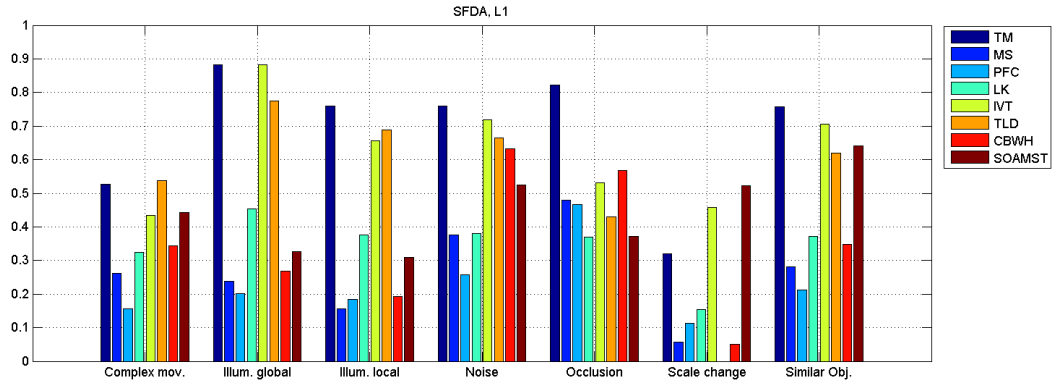


Figure 4.9: L1 SFDA result for all the individual trackers

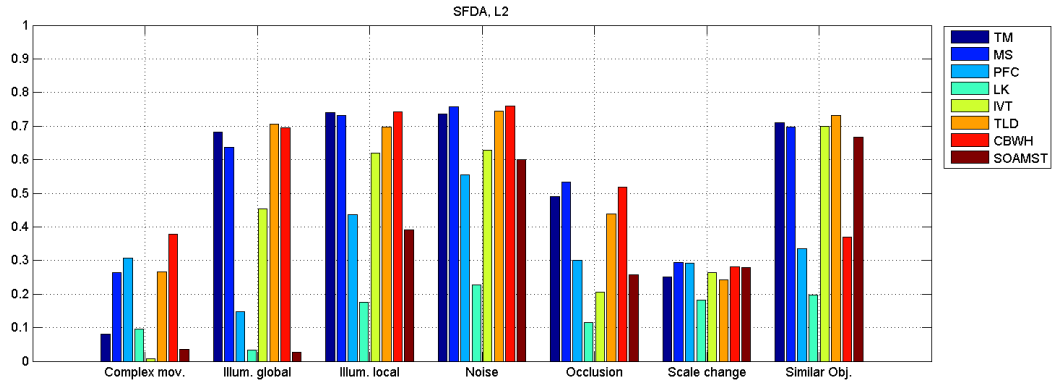


Figure 4.10: L2 SFDA result for all the individual trackers

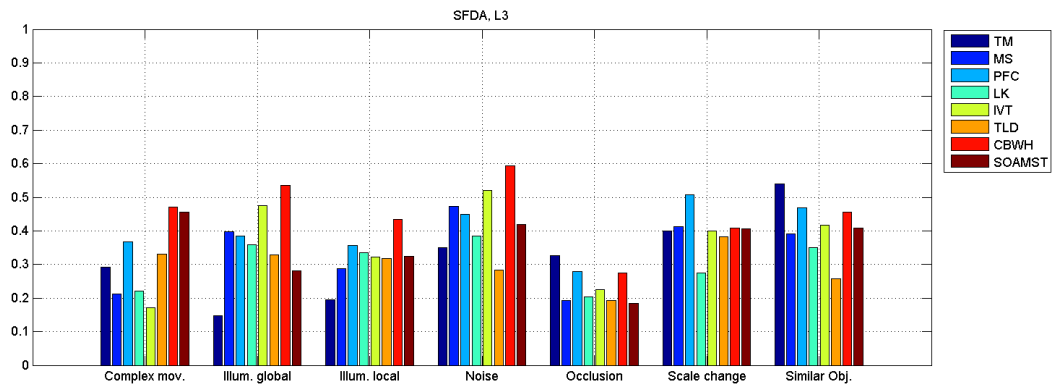


Figure 4.11: L3 SFDA result for all the individual trackers

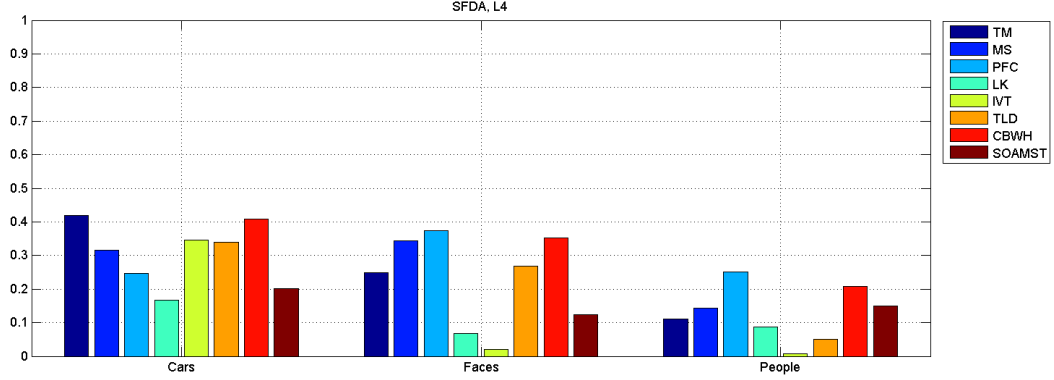


Figure 4.12: L4 SFDA result for all the individual trackers

The best results for the L1 sequences correspond with the algorithms TM, IVT and TLD. In the *scale change* category, the SOAMST and IVT algorithms present results significantly better than the others. The PFC algorithm presents the worse results in most of the subcategories of this set of sequences. TM presents the best results due to the simplicity of the sequences. As there are not appearance changes throughout the sequence, the initial template is the best possible model of the target. As there is no appearance changes, the IVT and TLD learning is successful independently of the frames in which the model is taken.

For the L2 sequences, the results obtained by the TM, MS, TLD and CBWH algorithms are the best in most cases, except in the subcategory of *similar objects* where the CBWH algorithm performs worse than the others. For the *scale changes subcategory*, the resulting scores are similar for the 8 trackers. Note that the results of all the algorithms in the category of *scale change* are always under 0.3. The LK tracker performs worse in most of the subcategories. This set of sequences is still relatively simple, so that the basic algorithms, TM and MS, work well due to the features presented in their definitions. The poor performance of the SOAMST algorithm compared with the MS one is surprising, being SOAMST a MS improvement. When considering more complex situations, the SOAMST algorithm believes there are changes in the sequence that does not happen and therefore fails.

In the L3 category, the results obtained by different trackers are similar in all subcategories. The CBWH tracker obtains the best score in *complex movement*, *global illumination*, *local illumination* and *noise* subcategories; the TM tracker obtains the best score in *occlusion* and *similar objects* subcategories; and the PFC tracker obtains the best score in *scale change* subcategory; but none of them stands out, in general terms, on the other trackers. The CBWH tracker stands above the rest as it achieves

reducing the effect of the background in the initialization frame. The remaining algorithms reduce their performance due to the increased difficulty of the sequences.

Finally, for the L4 sequences, the obtained scores are generally lower and worse than those obtained in the other 3 sets of sequences. The IVT algorithm presents the worse results in *faces* and *people* categories. The remaining trackers obtain low and similar results. As these sequences are not classified as in the above categories, there is less information to extract from the results.

## 4.4 Conclusions

As can be seen in the results, none of the trackers performs well in all categories and subcategories. The classical algorithms (TM, MS, PFC and LK) have limitations on certain types of sequences. Modern algorithms focus on specific problems (e.g., scale and orientation for SOAMST) or still have limitations in multiple types of sequences. That is why to study trackers combinations is interesting in order to achieve better overall results. The next chapter presents different types of fusions and their results. The aim of these fusions is to design a tracking approach that works well in most situations (sequences).



## Chapter 5

# Fusion

### 5.1 Introduction

In this chapter, some fusion methods are described and evaluated. After their definitions, tracking results on the dataset are presented and analyzed. Finally the results between individual trackers and fusions are compared, adding some final conclusions.

### 5.2 Fusion Methods

This section presents and explains the fusion methods used to combine the output of the individual trackers: mean, median and majority voting. The implemented fusion methods have been selected based on their simplicity and independence of individual algorithms (i.e., only using its outputs - bounding boxes).

Fusions considered use only the resulting bounding box of each of the individual trackers. For each frame, the bounding boxes resulting from the processing of each single tracking algorithm is extracted, and then the corresponding fusion is performed. Only the resulting bounding box from each tracker is used for the fusion, no matter what kind of single tracker has been used. Any tracker can be used for these types of fusion.

Furthermore, by using such simple methods, the computational cost of fusion is much less than the computational cost of individual trackers. If all the individual trackers are successful parallelized, the computing time of executing all the trackers and the subsequent fusion would be approximately equal to the time of the individual tracker with the greater computational time.

Figure 5.1 shows the block diagram of the fusion system.

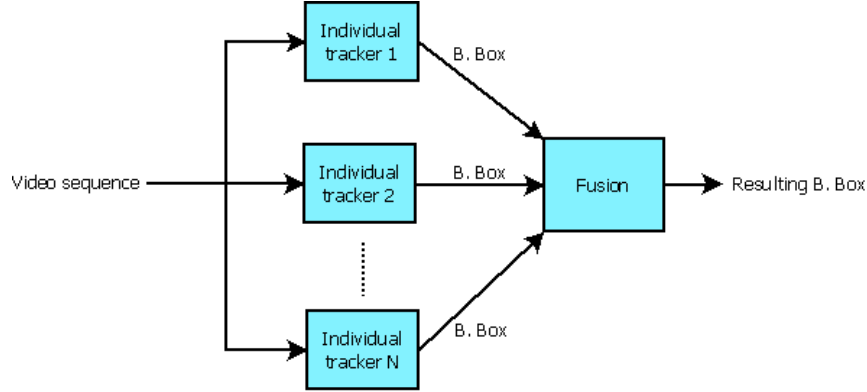


Figure 5.1: Fusion Block diagram

A general restriction has been added to all fusion methods: for a given frame, only the individual trackers whose resulting bounding box satisfies the following restrictions are considered for the fusion:

- Both height and width bounding box values must be different from zero.
- Both  $x_0$  and  $y_0$  values, corresponding to the center of the bounding box, must be defined values. This is because, in certain cases, some trackers return a NaN value when they suffer any malfunction.

### 5.2.1 Mean

The first considered fusion method is based on calculating the mean. Starting from all available resulting bounding boxes, the mean of the center coordinates of the bounding boxes ( $x_0$ ,  $y_0$ ), and of the height and width of the bounding boxes are calculated. These values are rounded to the nearest pixel value. In this way the new values that define the bounding box resulting from the fusion are obtained.

### 5.2.2 Median

This fusion method is very similar to media fusing method, except that in this case the operation performed is the median: the median of the center coordinates of the bounding boxes ( $x_0$ ,  $y_0$ ), and of the height and width of the bounding boxes are calculated. As in the previous fusion method, the values are rounded to the nearest pixel value.

### 5.2.3 Majority voting

Majority voting fusion is based on the selection of the resulting bounding box from the areas of the frame in which a minimum number of individual trackers coincide in indicating that the object is present. For a  $\geq N$  majority voting, the fusion resulting bounding box corresponds to the rectangle which contains all the areas in which at least  $N$  trackers agree that the object is located in that area.

## 5.3 Fusion results

The figures presented in this section show the SFDA score of the ten fusion methods considered: mean, median and 8 majority voting ( $N$  from  $\geq 1$  to  $\geq 8$ ). Majority voting  $\geq 1$  corresponds to logical OR, and majority voting  $\geq 8$  corresponds to logical AND. The results of the individual trackers have been added to facilitate the comparison between all the scores. Appendices B and C present the scores for all the metrics (see section 3.3.1).

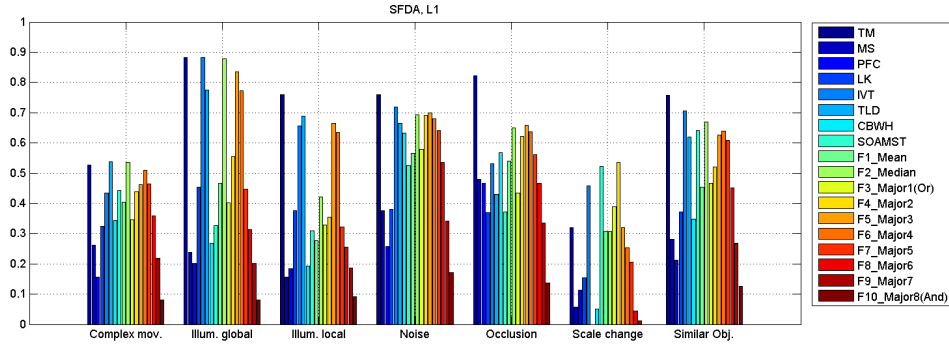


Figure 5.2: Fusion SFDA result of L1

For the L1 sequences, the TM results remain in most cases better than all others, including fusions. The score for TM is the best in *illumination local*, *noise*, *occlusion* and *similar objects* subcategories. This is because the template that is initialized in the TM model fits correctly the synthetic object on the different sequences. To track this type of synthetic objects, complex approaches are not needed. The best approach to track an object without appearance changes is template matching. When one of the trackers works considerably better than the others (as is the case of TM in L1) the fusion can not achieve those excellent results. For the complex *movement* and *illumination global*, the results of the median fusion are very similar to the TM ones. The most remarkable result is the *scale change* subcategory, where the score for

the majority voting ( $\geq 2$ ) fusion exceeds the SOAMST tracker score, which has been specially designed to address these kind of problems. When using a majority voting, the size of the final bounding box is variable. If multiple trackers indicate that the tracked object is centered in the same region but with slight variations, the result is a bounding box that better fits the size of the tracked object.

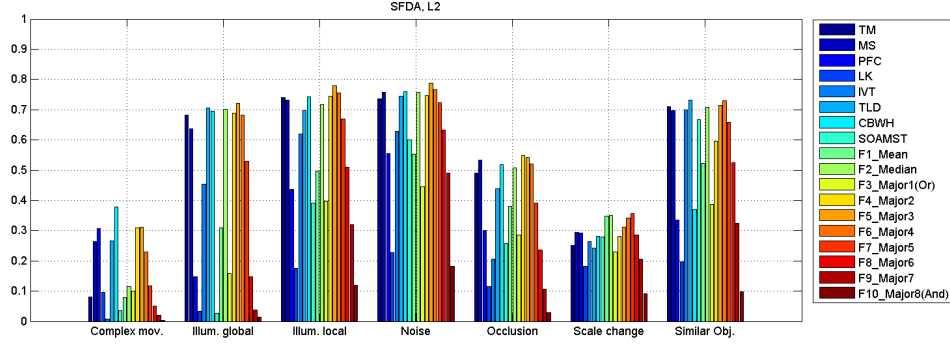


Figure 5.3: Fusion SFDA result of L2

In the case of the L2 categories, the results of median fusion and majority voting stand out. As in these sequences various trackers work relatively well in most cases, a fusion by majority voting is a good choice. The same happens in the case of median fusion, when either a majority tracks correctly the object or the erroneous values are positioned at the extremes of the ordination of values for calculating the median. For the *illumination global*, *illumination local*, *noise*, *occlusions* and *similar object* subcategories, the obtained results are similar to those of the best individual trackers. In the case of the *complex movement* subcategory, the result obtained by the CBWH tracker remains better than the rest of the individual trackers and fusions. The score obtained with some fusions for the *scale change* problem is significantly better than that obtained by the individual trackers. The justification of this result is the same than in the L1 scale change sequences.

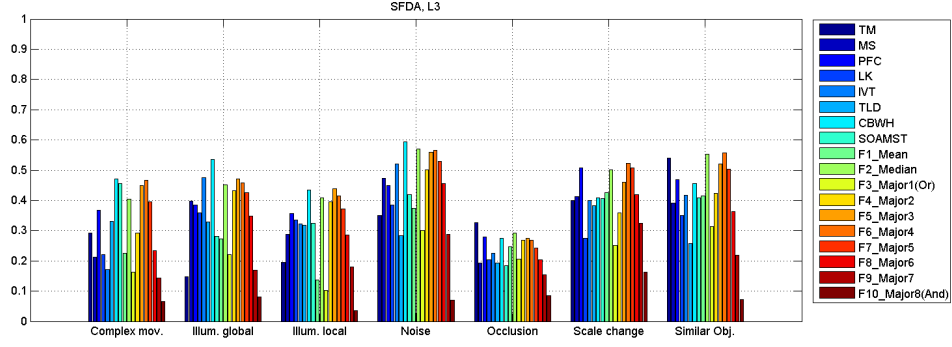


Figure 5.4: Fusion SFDA result of L3

Median and central majority voting ( $\geq 3$ ,  $\geq 4$  and  $\geq 5$ ) present similar results to the best individual trackers in L3 sequences. The best result (CBWH) for *illumination global* and *noise* subcategories has not been reached with the fusions. Fusions performance is similar to that described for the L2 sequences.

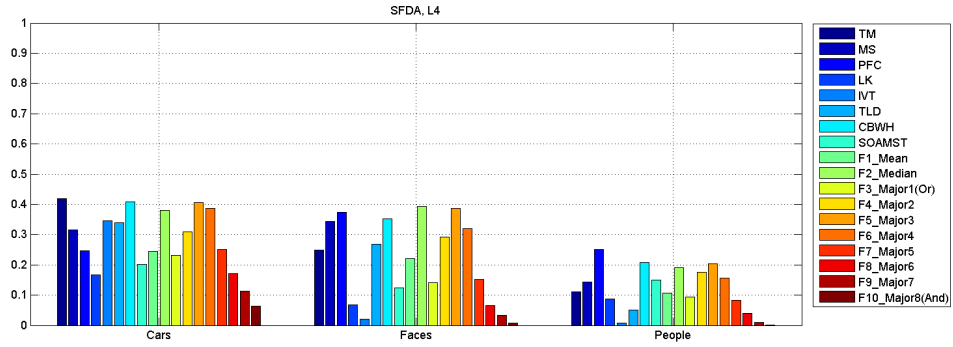


Figure 5.5: Fusion SFDA result of L4

Finally, for L4 sequences, median and majority voting ( $\geq 3$  and  $\geq 4$ ) fusions present similar results to the best achieved by the individual trackers. Note that in these sequences the scores are generally quite low, due to the difficulty presented. When all the individual trackers do not work correctly (low scores), the result of the fusion is difficult to be better than the best of the individual trackers.

## 5.4 Comparative results between individual trackers and fusions

After analyzing the obtained results, the most interesting aspect of the studied fusions is the versatility of some of them. The tracking methods designed, as a combination of individual trackers, work well in most cases.

The tables contained in this section present the individual global scores for each metric and each tracker. These global scores are obtained as follows: for each metric and each tracker, all the dataset subcategories results are added. It is, 7 for the L1 category, 7 for the L2 category, 7 for the L3 category and 3 for the L4 category. The result is then normalized between 0 and 1, dividing the result by 24 (7+7+7+3). In this way, we obtain a single score that contains information of all subcategories which allows to know the overall performance of the tracker for all the sequences.

Table 5.1 presents the global scores of both individual trackers as fusions.

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
SFDA	0,481	0,372	0,319	0,246	0,424	0,429	0,429	0,348

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
SFDA	0,349	0,507	0,291	0,462	0,525	0,511	0,423	0,308	0,196	0,078

Table 5.1: global SFDA scores

To facilitate the results comparison, tables 5.2 and 5.3 show the difference and percentual difference, respectively, between the individual trackers and fusions. The differences between the global scores obtained for each fusion and the global scores obtained for each individual tracker algorithm are obtained by subtraction (fusion score - individual score). As both scores range from 0 to 1, the difference also ranges from 0 to 1 (The result in the difference tables is also multiplied by 100 to get the percentage). The percentual differences between the global scores obtained for each fusion and the global scores obtained for each individual tracker algorithm are obtained as presented in equation 5.1.

$$Percentual\ difference = \frac{fusion\ score - individual\ score}{individual\ score} \quad (5.1)$$

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-13,20	-2,33	3,01	10,26	-7,48	-8,01	-7,99	0,09
F2_Median	2,57	13,45	18,79	26,03	8,29	7,77	7,78	15,86
F3_Major1	-19,02	-8,14	-2,80	4,44	-13,29	-13,82	-13,81	-5,73
F4_Major2	-1,93	8,94	14,29	21,53	3,79	3,26	3,28	11,36
F5_Major3	<b>4,41</b>	<b>15,28</b>	<b>20,63</b>	<b>27,87</b>	<b>10,13</b>	<b>9,60</b>	<b>9,62</b>	<b>17,70</b>
F6_Major4	3,02	13,90	19,24	26,48	8,75	8,22	8,23	16,31
F7_Major5	-5,77	5,10	10,45	17,69	-0,05	-0,58	-0,56	7,52
F8_Major6	-17,27	-6,40	-1,05	6,19	-11,55	-12,08	-12,06	-3,98
F9_Major7	-28,50	-17,63	-12,28	-5,04	-22,78	-23,31	-23,29	-15,21
F10_Major8	-40,30	-29,43	-24,08	-16,84	-34,58	-35,11	-35,09	-27,01

Table 5.2: Difference (percentage) SFDA global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-27,45	-6,26	9,46	41,63	-17,65	-18,67	-18,63	0,25
F2_Median	5,35	36,12	58,94	105,67	19,58	18,10	18,15	45,57
F3_Major1	-39,54	-21,88	-8,78	18,03	-31,38	-32,22	-32,19	-16,46
F4_Major2	-4,01	24,02	44,82	87,39	8,95	7,61	7,65	32,63
F5_Major3	<b>9,17</b>	<b>41,06</b>	<b>64,71</b>	<b>113,13</b>	<b>23,91</b>	<b>22,39</b>	<b>22,44</b>	<b>50,85</b>
F6_Major4	6,29	37,33	60,36	107,50	20,64	19,15	19,20	46,87
F7_Major5	-12,00	13,71	32,77	71,80	-0,11	-1,34	-1,30	21,60
F8_Major6	-35,91	-17,18	-3,30	25,13	-27,25	-28,15	-28,12	-11,43
F9_Major7	-59,26	-47,36	-38,53	-20,46	-53,76	-54,33	-54,31	-43,70
F10_Major8	-83,79	-79,06	-75,55	-68,36	-81,61	-81,83	-81,82	-77,61

Table 5.3: Percentual difference (percentage) SFDA global score between fusion trackers and individual algorithms trackers

As shown in the tables, both the median fusion and the majority voting ( $\geq 3$  and  $\geq 4$ ) fusions have better global SFDA results than any of the individual trackers.

Finally, tables 5.4 and 5.5 present the differences and percentual differences (see eq. 5.1) between the global scores obtained for each fusion and the best one of global scores obtained for all the individual tracker algorithm.

	SFDA	ATA	ATEinv	AUCinv	PixelOv	CTM	TC	CoTPSinv
F1_Mean	-13,20	-13,31	-17,17	-11,16	-11,23	-13,31	-9,37	-9,859
F2_Median	2,57	2,60	-3,62	7,90	7,83	2,60	5,01	<b>6,187</b>
F3_Major1	-19,02	-19,16	-40,95	-21,17	-21,10	-19,16	-23,36	-28,654
F4_Major2	-1,93	-1,94	-21,27	-0,54	-0,54	-1,94	3,98	-6,753
F5_Major3	<b>4,41</b>	<b>4,45</b>	-8,23	<b>8,79</b>	<b>8,79</b>	<b>4,45</b>	<b>11,67</b>	4,226
F6_Major4	3,02	3,05	1,63	7,62	7,61	3,05	8,20	4,419
F7_Major5	-5,77	-5,81	7,57	-2,85	-2,91	-5,81	-1,53	-2,200
F8_Major6	-17,27	-17,40	10,75	-16,84	-16,98	-17,40	-14,99	-14,565
F9_Major7	-28,50	-28,72	15,57	-30,15	-30,36	-28,72	-28,61	-25,669
F10_Major8	-40,30	-40,62	<b>20,72</b>	-43,62	-43,91	-40,62	-53,88	-37,1192627

Table 5.4: Difference (percentage) global score between fusion trackers and best individual algorithms trackers

	SFDA	ATA	ATEinv	AUCinv	PixelOv	CTM	TC	CoTPSinv
F1_Mean	-27,45	-27,45	-24,11	-21,54	-21,49	-27,45	-12,82	-16,59
F2_Median	5,35	5,37	-5,08	15,26	14,99	5,37	6,86	<b>10,41</b>
F3_Major1	-39,54	-39,52	-57,51	-40,86	-40,38	-39,52	-31,97	-48,20
F4_Major2	-4,01	-4,01	-29,88	-1,05	-1,03	-4,01	5,44	-11,36
F5_Major3	<b>9,17</b>	<b>9,19</b>	-11,55	<b>16,97</b>	<b>16,82</b>	<b>9,19</b>	<b>15,97</b>	7,11
F6_Major4	6,29	6,30	2,29	14,71	14,56	6,30	11,22	7,43
F7_Major5	-12,00	-11,99	10,63	-5,51	-5,57	-11,99	-2,09	-3,70
F8_Major6	-35,91	-35,89	15,10	-32,52	-32,50	-35,89	-20,52	-24,50
F9_Major7	-59,26	-59,25	21,86	-58,20	-58,11	-59,25	-39,16	-43,18
F10_Major8	-83,79	-83,79	<b>29,10</b>	-84,20	-84,04	-83,79	-73,74	-62,44

Table 5.5: Percentual difference (percentage) global score between fusion trackers and best individual algorithms trackers

In this latter result, only a majority voting  $\geq 4$  (corresponding with a majority voting  $\geq 50\%$  of the votes) gets higher score than any of the individual trackers for any of the metrics presented in section 3.3.1. This improvement in most of the metrics is limited, but must be considered that the fusion is being compared with the best tracking of the eight individual algorithms. The median fusion also gets better results than all the individual trackers except for the ATE metric.

Another notable result is that when the minimum number for the majority voting is increased, the score of the ATE metric increases, because, being more restrictive, the number of false positives decreases.



## 5.5 Results extracted from the the SoA fusion algoritms

This section presents a comparison between the results presented in the papers of each fusion algorithm from the state of the art and the results of the fusions presented in this chapter.

[36] initially uses 4 videos to get cualitative results of its tracker performance (accuracy). The mean center location error is 11 pixels. After that, they constructed another set of 7 sequences with illunimation changes, occlusions, noise, etc., giving in this case a mean center location error of 12.28 pixels. In [35] the same result is presented with a different set of 7 videos, with a score of 9.14 pixels of mean center location error. In [31] the same result is presented with a different set of 6 videos, with a score of 11.16 pixels of mean center location error. The results presented are very good, but the number of sequences is very low and has been individually selected by the authors. Our dataset contains a larger set of videos. Moreover, the metrics used are more complete than the one used to give these results.

The results presented in [33] are based on three selected sequences. The metrics presented are the percentages of “dropouts” and “errors”. The results are 8.2, 14.8 and 6.1 for the dropouts percentages. For the error percentage measure, the values are 5.0, 16.2 and 1.1. As in the two previous references, using three sequences is not enough information to get overall results or conclusions. The metrics used neither help us understand the characteristics of the tracker.

The results presented in [34] are more complete than those presented by the previous references. The results are based on Caremedia[53] (13 sequences) and Caviar[47] (79 videos) datasets. The proposed fusion improves a 28.89% and a 26.63% over the best individual F1 score of the trackers, depending on the set of trackers used (a set of 5 or 10 trackers). In our case, fusions that improve the best of every single tracker are also achieved (see table 5.5). To compare our results with the results of [34], both should be obtained with the same set of sequences.

The Caremedia[53] dataset (13 sequences) is also used in [32]. The score used is the F1 score. The performance of the proposed fusion in the paper is better than the average fusion score of all the individual tracker used, but it is worse than the best individual tracker score. As mentioned in the previous paragraph, to compare our results with the results of [32], both should be obtained with the same set of sequences.

Most of the sequences used for the evaluation in [16] have been downloaded from *YouTube*. The proposed tracker tracks correctly (100% score) most of the used sequences. In this evaluation, a frame is countered as correctly tracked if overlap score

is  $>0.5$  (see subsection 3.3.1.4).

## 5.6 Conclusions

After experimenting with some simple fusions, some combinations of trackers with better results than the 8 independent tracking algorithms have been obtained. The best results have been obtained with the median fusion and with the majority voting (with around 50% of the votes).

The main problem of the individual trackers is that they usually work well only for certain environments, posing great difficulties for others. With the trackers combinations, good results have been achieved in most environments considered, solving the problem mentioned above.

The considered fusion types perform poorly when there is a tracker which functions considerably better than the rest, as the combination result is not able to achieve these good results.

## Chapter 6

# Conclusions and future work

### 6.1 Introduction

In this chapter, the conclusions of the developed work and some future work lines are presented.

### 6.2 Conclusions

The work presented in this document is focused on video object tracking. This field of study is one of the most popular in Computer Vision, so there is abundant literature, algorithms, metrics, datasets, etc. about this subject.

There are multiple datasets for video object tracking. Depending on the objective of your work, there are some appropriate datasets. One possibility is to combine these datasets to form a new one that suits your needs, as is the case of the dataset used in this work. About the metrics, there are also multiple possibilities for evaluating a video object tracker. In general, these metrics are highly correlated, as all attempt to measure how well the target object is tracked. The main differences between the metrics are based on the penalties that are attributed to the errors (false positives, false negatives, target loss ...).

About tracking algorithms, there are many publications that present their own tracking algorithms, and many others which try to improve some aspect or limitation of existing algorithms. As observed in the results of individual trackers, all tracking algorithms have limitations in certain scenarios, and only work well in those scenarios for which they were designed. Note that even classical algorithms as the TM one present, in some scenarios, better performance than modern algorithms.

With the (simple) fusions performed, more versatile trackers have been obtained,

which are able to function reasonably well in most situations (covered by the selected dataset), overcoming the problem of specialization observed in individual trackers.

### 6.3 Future Work

The work described in this document analyzes several algorithms for tracking from the state of the art and presents some methods to combine them efficiently. Despite this, a tracker that performs properly in all possible situations has not yet been achieved, as there are problems which are not solved with the used algorithms. Moreover, there are new scenarios not covered in this Master Thesis. We identify some main areas for future work:

- With respect to individual trackers, there are two main possibilities:
  - ✧ Development of a new individual tracker trying to overcome the problems observed in the analyzed algorithms. Thanks to the evaluation framework, the results of the developed algorithms can be easily compared with the reference algorithms results.
  - ✧ Inclusion of additional algorithms designed by other authors. New algorithms can be analyzed and their results can be compared with the previous algorithms and be incorporated in the fusion approaches.
- New fusion methods can be studied and evaluated, for example, by adding weights to the different algorithms depending on its accuracy. Another possibility is to add feedback to the system, so that the result of each frame can be used to adjust the analysis of the subsequent frames.
- The research can be extended to multitarget and multi-camera systems. To do this, new content sets that contain these types of videos should be chosen or created. Also the metrics used should be reconsidered.

# Bibliography

- [1] M. Lozano, “Evaluación comparativa de algoritmos de seguimiento de objetos (tracking),” Master’s thesis, Universidad Autónoma de Madrid, 2012.
- [2] A. S. Jalal and V. Singh, “The state-of-the-art in visual object tracking,” *Informatica*, vol. 36, no. 3, pp. 227–248, 2012.
- [3] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *ACM Comput. Surv.*, vol. 38, Dec. 2006.
- [4] E. Maggio and A. Cavallaro, *Video Tracking: Theory and Practice*. Wiley, 2011.
- [5] I. J. Cox, “A review of statistical data association techniques for motion correspondence,” *International Journal of Computer Vision*, vol. 10, pp. 53–66, 1993.
- [6] Y. Bar-Shalom, *Tracking and data association*. San Diego, CA, USA: Academic Press Professional, Inc., 1987.
- [7] Y. Bar-Shalom and E. Tse, “Tracking in a cluttered environment with probabilistic data association,” *Automatica*, vol. 11, pp. 451–460, Sept. 1975.
- [8] S. S. Blackman, *Multiple-target tracking with radar applications*. Artech House radar library, Norwood, Mass. Artech House, 1986.
- [9] Y. Bar-Shalom, “Tracking methods in a multitarget environment,” *In trans. on Automatic Control*, vol. 23, no. 4, pp. 618–626, 1978.
- [10] G. Pulford, “Taxonomy of multiple target tracking methods,” *In proc of. Radar, Sonar and Navigation*, vol. 152, no. 5, pp. 291–304, 2005.
- [11] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, “Multicamera people tracking with a probabilistic occupancy map,” *In trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.
- [12] S. Khan and M. Shah, “Tracking multiple occluding people by localizing on multiple scene planes,” *In trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 505–519, 2009.
- [13] J. Black, T. Ellis, and P. Rosin, “Multi view image surveillance and tracking,” in *In proc. of Workshop on Motion and Video Computing*, pp. 169–174, 2002.

- [14] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in *In proc. of Workshop on Visual Surveillance*, pp. 3–10, 2000.
- [15] M. Taj and A. Cavallaro, "Distributed and decentralized multi-camera tracking," *Signal Processing Magazine*, vol. 28, no. 3, 2011.
- [16] M. Heber, M. Godec, M. R  ther, P. M. Roth, and H. Bischof, "Segmentation-based tracking by support fusion," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 573–586, 2013.
- [17] EventVideo, "Deliverable 5.3v1, eventvideo test sequences, ground-truth and evaluation methodology," 2012.
- [18] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *In trans. on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.
- [19] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley Publishing, 2009.
- [20] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *In trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.
- [21] K. Nummiaro, E. Koller-Meier, and L. V. Gool, "An adaptive colour-based particle filter," *In proc. of Image and Vision Computing*, vol. 21, no. 1, pp. 99–110, 2002.
- [22] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, pp. 221–255, March 2004.
- [23] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, pp. 125–141, May 2008.
- [24] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *In trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2011.
- [25] J. Ning, L. Zhang, D. Zhang, and C. Wu, "Robust mean-shift tracking with corrected background-weighted histogram," *Computer Vision, IET*, vol. 6, no. 1, pp. 62–69, 2012.
- [26] J. Ning, L. Zhang, D. Zhang, and C. Wu, "Scale and orientation adaptive mean shift tracking," *Computer Vision, IET*, vol. 6, no. 1, pp. 52–61, 2012.
- [27] B. Stenger, T. Woodley, and R. Cipolla, "Learning to track with multiple observers.," in *In proc. of Computer Vision and Pattern Recognition*, pp. 2647–2654, IEEE, 2009.
- [28] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
- [29] I. Leichter, M. Lindenbaum, and E. Rivlin, "A general framework for combining visual trackers, the black boxes approach," *Journal of Computer Vision*, vol. 67, pp. 343–363, 2006.

- [30] S. Avidan, “Ensemble tracking,” *In trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261–271, 2007.
- [31] B. Zhong, H. Yao, S. Chen, R. Ji, X. Yuan, S. Liu, and W. Gao, “Visual tracking via weakly supervised learning from multiple imperfect oracles,” in *In proc. of Computer Vision and Pattern Recognition*, pp. 1323–1330, 2010.
- [32] W. Liu and A. G. Hauptmann, “A crowdsourcing approach to tracker fusion,” tech. rep., Carnegie Mellon University, 2011.
- [33] B. McCane, B. Galvin, and K. Novins, “Algorithmic fusion for more robust feature tracking,” *Journal of Computer Vision*, vol. 49, pp. 79–89, 2002.
- [34] L. Zhang, Y. Gao, A. Hauptmann, R. Ji, G. Ding, and B. Super, “Symbiotic black-box tracker,” in *In proc. of international conference on Advances in Multimedia Modeling, MMM’12*, (Berlin, Heidelberg), pp. 126–137, Springer-Verlag, 2012.
- [35] J. Kwon and K. M. Lee, “Visual tracking decomposition,” in *In proc. of Computer Vision and Pattern Recognition*, pp. 1269–1276, 2010.
- [36] J. Kwon and K. M. Lee, “Tracking by sampling trackers,” in *In proc. of International Conference on Computer Vision*, pp. 1195–1202, 2011.
- [37] H. Wang, C. Liu, L. Xu, M. Tang, and X. Wu, “Multiple feature fusion for tracking of moving objects in video surveillance,” in *In conf. on Computational Intelligence and Security*, vol. 1, pp. 554–559, 2008.
- [38] C. O. Conaire, N. E. O’Connor, and A. F. Smeaton, “Detector adaptation by maximising agreement between independent data sources,” in *In proc. of Computer Vision and Pattern Recognition*, pp. 1–6, 2007.
- [39] J. C. SanMiguel and J. M. Martinez, “Shadow detection in video surveillance by maximizing agreement between independent detectors,” in *In trans. on Image Processing*, pp. 1141–1144, 2009.
- [40] MIT, “traffic data set, <http://www.ee.cuhk.edu.hk/~xgwang/mittraffic.html>, last accessed, 24 may 2012,”
- [41] I. für Algorithmen und Kognitive Systeme, “Cars dataset, <http://i21www.ira.uka.de/image-sequences/>, last accessed, 24 may 2012,”
- [42] T. 2009, “Event detection dataset, <http://trecvid.nist.gov/trecvid.data.html>, last accessed, 24 may 2012.,”
- [43] S. Birchfield, “Elliptical head tracking using intensity gradients and color histograms, <http://www.ces.clemson.edu/stb/research/headtracker/>,”
- [44] R. Vezzani and R. Cucchiara, “Video surveillance online repository (visor): an integrated framework,” *Multimedia Tools and Applications*, vol. 50, no. 2, pp. 359–380, 2010.
- [45] SPEVI, “surveillance performance evaluation initiative, <http://www.eecs.qmul.ac.uk/andrea/spevi.html>, last accessed, 24 may 2012,”

- [46] PETS, “Ieee int. workshop perform. eval. track. surveill.,last accessed, 24 may 2012.”
- [47] CAVIAR, “Context aware vision using image-based active recognition, <http://homepages.inf.ed.ac.uk/rbf/caviar/>, last accessed, 24 may 2012.”
- [48] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, “Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol,” *In trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.
- [49] D. M. Chu, A. W. M., and Smeulders, “Thirteen hard cases in visual tracking,” in *In proc. of Advanced Video and Signal Based Surveillance*, pp. 103–110, 2010.
- [50] T. Nawaz and A. Cavallaro, “Pft: A protocol for evaluating video trackers,” in *In proc. of International Conference on Image Processing*, pp. 2325–2328, 2011.
- [51] F. Yin, D. Makris, and S. A. Velastin, “Performance evaluation of object tracking algorithms,” in *In proc. of Performance Evaluation of Tracking and Surveillance*, 2007.
- [52] T. Nawaz and A. Cavallaro, “A protocol for evaluating video trackers under real-world conditions,” *In trans. on Image Processing*, 2012.
- [53] S. Stevens, D. Chen, H. Wactlar, A. Hauptmann, M. Christel, and A. Bharucha, “Automatic collection, analysis, access and archiving of psycho/social behavior by individuals and groups,” in *ACMWorkshop on Continuous Archival and Retrival of Personal Experiences*, 2006.
- [54] S. Munder and D. M. Gavrilu, “An experimental study on pedestrian classification,” *In trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1863–1868, 2006.



# List of Abbreviations

TM	Template Matching tracker
MS	Mean-Shift tracker
PFC	Particle Filter-based Colour tracker
LK	Lucas-Kanade tracker
IVT	Incremental learning for robust Visual Tracking tracker
TLD	Tracking Learning Detection tracker
CBWH	Corrected Background-Weighted Histogram mean-shift tracker
SOAMST	Scale and Orientation Adaptive Mean-Shift tracker
SFDA	Sequence Frame Detection Accuracy
ATA	Average Tracking Accuracy
ATE	Average Tracking Error
AUC	Area Under the lost track ratio Curve
CTM	Closeness of Track (Mean)
CTD	Closeness of Track (Deviation)
TC	Track Completeness
CoTPS	Combined Tracking Performance Score



## Appendix A

# Video object tracking datasets<sup>1</sup>

This appendix presents some publicly available state of the art datasets. The SOVTds presentend in 3.2 is composed with sequences of these datasets.

### A.1 SPEVI

The Surveillance Performance EValuation Initiative (SPEVI) [45] is a set of links of publicly available datasets for researches. The videos can be used for testing and evaluating video tracking algorithms for surveillance-related applications. Two datasets are especially interesting regarding the tracking evaluation and they are described as follows.

#### A.1.1 Single Face Dataset

This is a dataset for single person/face visual detection and tracking. The sequences include different illumination conditions and resolutions.

- Number of sequences: 5 sequences, 3018 frames.
- Format: individual JPEG images.
- Tracking ground-truth available: yes.

#### A.1.2 Multiple Face Dataset

This is a dataset for multiple people/faces visual detection and tracking. The sequences (same scenario) contain 4 targets which repeatedly occlude each other while appearing and disappearing from the field of view of the camera.

---

<sup>1</sup>This appendix has been extracted from [17].

- Number of sequences: 3 sequences, 2769 frames.
- Format: individual JPEG images.
- Tracking ground-truth available: yes.



Figure A.1: Sample frames for the SPEVI dataset (top: single object, down: multiple object)

## A.2 ETISEO

ETISEO [54] is a video understanding evaluation project that contains indoor and outdoor scenes, corridors, streets, building entries, subway, etc. This content set also mix different types of sensors and complexity levels.

- Number of sequences: 86 sequences.
- Tracking ground-truth available: yes.



Figure A.2: Sample frames for the ETISEO dataset

### A.3 PETS

PETS [46] is the most extended database nowadays. A new database is released each year since 2000, along with a different challenge proposed. With the algorithms provided researchers can test or develop new algorithms. The best ones are presented in the conference held each year. Since the amount of data is extensive and cover real situations, these databases are by far the most used and are almost considered a de facto standard. Despite this, it is important to say that the PETS databases are not ideal. One of its disadvantages is the fact that since PETS became a surveillance project, the challenges proposed are focused on high level applications of that field, leaving aside the tracking approach. Therefore, some important issues (such as illumination or target scale changes) are not considered.

#### A.3.1 PETS2000

Outdoor people and vehicle tracking (single camera).

- Number of sequences: 1 set of training and test sequence.
- Training sequence: 3672 frames.
- Test sequence: 1452 frames.
- Formats: MJPEG movies and JPEG frames.
- Tracking ground-truth available: no.

#### A.3.2 PETS 2001

Outdoor people and vehicle tracking (two synchronized views; includes omnidirectional and moving camera). Challenging in terms of significant lighting variation, occlusion, scene activity and use of multi-view data.

Number of sequences: 5 sets of training and test sequences

Training sequences: 1st) 3064 frames. 2nd) 2989 frames. 3rd) 5563 frames. 4th) 6789 frames. 5th) 2866 frames.

Test sequences: 1st) 2688 frames. 2nd) 2823 frames . 3rd) 5336 frames. 4th) 5010 frames. 5th) 2867 frames.

Formats (for each set): MJPEG movies and JPEG frames.

Tracking ground-truth available: no.

### A.3.3 PETS 2006

Multicamera person and baggage detection in a train station. Scenarios of increasing complexity, captured using multiple sensors.

- Number of sequences: 7 sets with 4 cameras each.
- Formats (for each set): MJPEG movies and JPEG frames.
- Tracking ground-truth available: no.



Figure A.3: Sample frames for the PETS2006 dataset

### A.3.4 PETS 2007

Multicamera setup containing the following scenarios: loitering; attended luggage removal (theft) and unattended luggage with increasing scene complexity.

- Number of sequences: 1 training set + 9 testing sets.
- Formats (for each set): JPEG frames.
- Tracking ground-truth available: no.

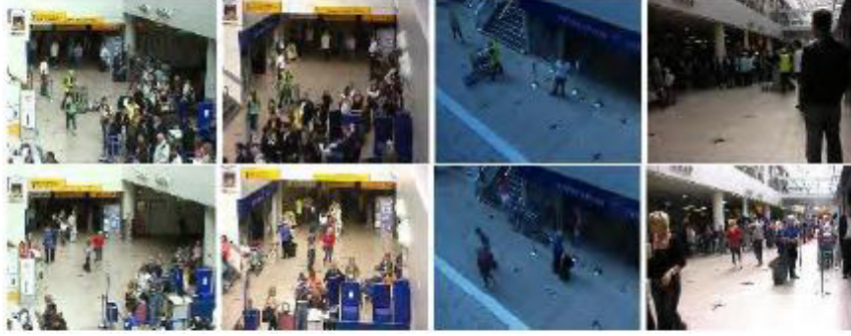


Figure A.4: Sample frames for the PETS2007 dataset

### A.3.5 PETS 2010

Multicamera setup containing different crowd activities (these datasets are the same as used for PETS2009).

- Number of sequences: 1 training set + 3 testing sets.



Figure A.5: Sample frames for the PETS2010 dataset

## A.4 CAVIAR

The main objective of CAVIAR [47] is to address the scientific question: Can rich local image descriptions from foveal and other image sensors, selected by a hierarchical visual attention process and guided and processed using task, scene, function and object contextual knowledge improve image-based recognition processes. Several methods were researched in order to address this question, including different areas, and the results were integrated in a closed-loop object and situation recognition system. This dataset includes sequences of people walking alone, meeting with others,

window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place. All video clips were filmed with a wide angle camera lens, and some scenarios were recorded with two different points of view (synchronized frame by frame).

- Number of sequences: INRIA (1st set): 6 sequences, Shopping Center in Portugal (2nd set): 11 sequences, 6 different scenarios.
- Formats (for both sets): MJPEG movies, JPEG frames, XML ground-truth.
- Tracking ground-truth available: yes.



Figure A.6: Sample frames for the CAVIAR dataset

## A.5 VISOR

The Video Surveillance Online Repository is an extensive database containing a large set of multimedia data and the corresponding annotations. The repository has been conceived as a support tool for different research projects [44]. Some videos are available publicly; however, most of them are restricted and can only be viewed after a registration. The videos in the database cover a wide range of scenarios and situations, including (but not limited to) videos for human action recognition, outdoor videos for face detection, indoor videos for people tracking with occlusions, videos for human recognition, videos for vehicles detection and traffic surveillance. This dataset includes several videos with a wide range of occlusions caused by objects or people in the scene. All of them include base annotations and some also include automatic annotations.



- Number of sequences: 6 sequences.
- Format: MJPEG movies.
- Tracking ground-truth available: no.



Figure A.7: Sample frames for the VISOR dataset

## A.6 iLids

The Imagery Library for Intelligent Detection Systems (i-Lids) bag and vehicle detection challenge was included in the 2007 AVSS Conference [45]. This dataset includes several sequences for two separate tasks: first, an abandoned baggage scenario and second, a parked vehicle scenario.

- Number of sequences: 7 sequences (3 for Task 1, 4 for Task 2).
- Format: JPEG images, 8-bit color MOV, XML for ground-truth.
- Tracking ground-truth available: no.



Figure A.8: Sample frames for the i-LIDS dataset

## A.7 Clemson dataset

Included in an elliptical head tracking project by Stan Birchfield there is a series of videos very interesting for head tracking. The sequences include issues such as occlusion, rotation, translation, clutter in the scene, change in the target's size, etc. The tracker as well as the sequences can be found at the web [43]. This dataset includes several sequences for head tracking with different targets.

- The videos include some of the most important issues for tracking algorithms.
- Number of sequences: 16 short sequences (1350 frames in total).
- Format: BMP images.
- Ground-truth available: yes.



Figure A.9: Sample frames for the CLEMSON dataset

## A.8 MIT Traffic Dataset

MIT traffic dataset is for research on activity analysis and crowded scenes. It includes a traffic video sequence of 90 minutes long recorded by a stationary camera. The size of the scene is 720 by 480. More information regarding this work can be found in [40]. This dataset includes several clips regarding traffic. It contains a representation of most of the issues previously described, making this a very interesting dataset.

- Number of sequences: 1 sequence, 165880 frames divided in 20 clips.



Figure A.10: Sample frames for the MIT traffic dataset



## Appendix B

# Trackers and fusion results: Data tables

This appendix presents the obtained fusion results with tables. Four figures (one for each dataset category) are presented for each one of the nine metrics.

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.53	0.08	0.29	0.88	0.68	0.15	0.76	0.74	0.20	0.76	0.74	0.35	0.82	0.49	0.33	0.32	0.25	0.40	0.76	0.71	0.54	0.42	0.25	0.11
	0.53	0.08	0.30	0.89	0.68	0.15	0.77	0.74	0.20	0.77	0.74	0.35	0.83	0.49	0.33	0.32	0.25	0.40	0.77	0.71	0.54	0.42	0.25	0.11
	0.62	0.10	0.38	0.94	0.76	0.20	0.87	0.80	0.23	0.87	0.79	0.44	0.89	0.54	0.39	0.95	0.31	0.43	0.87	0.80	0.63	0.46	0.32	0.14
ATEnv																								
AUCInv	0.71	0.09	0.37	0.84	0.72	0.23	0.84	0.78	0.21	0.84	0.78	0.43	0.77	0.51	0.36	0.33	0.25	0.41	0.84	0.74	0.57	0.42	0.26	0.13
PixelOv	0.71	0.09	0.38	0.85	0.73	0.23	0.85	0.79	0.21	0.85	0.78	0.43	0.78	0.51	0.36	0.34	0.26	0.41	0.85	0.74	0.57	0.42	0.27	0.13
CTM	0.53	0.08	0.30	0.89	0.68	0.15	0.77	0.74	0.20	0.77	0.74	0.35	0.83	0.49	0.33	0.32	0.25	0.40	0.77	0.71	0.54	0.42	0.25	0.11
CTD	0.14	0.23	0.26	0.04	0.13	0.23	0.03	0.10	0.18	0.03	0.10	0.20	0.14	0.25	0.19	0.24	0.18	0.21	0.03	0.16	0.20	0.30	0.25	0.19
TC	0.77	0.12	0.52	1.00	1.00	0.30	1.00	1.00	0.30	1.00	1.00	0.59	0.97	0.69	0.52	0.56	0.45	0.73	1.00	0.92	0.78	0.63	0.46	0.20
CoTPSInv	0.75	0.21	0.58	0.84	0.72	0.46	0.84	0.78	0.32	0.84	0.78	0.57	0.78	0.62	0.47	0.32	0.40	0.46	0.84	0.79	0.60	0.58	0.44	0.28

Table B.1: Results for TM tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.26	0.26	0.21	0.24	0.64	0.40	0.16	0.73	0.29	0.38	0.76	0.47	0.48	0.53	0.19	0.06	0.29	0.41	0.28	0.70	0.39	0.32	0.34	0.14
	0.26	0.26	0.22	0.24	0.64	0.40	0.16	0.73	0.29	0.38	0.76	0.48	0.48	0.54	0.19	0.06	0.29	0.41	0.28	0.70	0.39	0.32	0.35	0.15
ATA	0.38	0.33	0.28	0.32	0.72	0.54	0.24	0.79	0.38	0.55	0.81	0.61	0.57	0.61	0.26	0.21	0.35	0.46	0.42	0.80	0.48	0.37	0.45	0.22
ATEInv	0.39	0.33	0.26	0.25	0.66	0.57	0.22	0.78	0.39	0.52	0.81	0.60	0.52	0.54	0.21	0.11	0.30	0.42	0.39	0.73	0.41	0.31	0.37	0.21
AUCInv	0.40	0.33	0.26	0.25	0.66	0.57	0.23	0.78	0.39	0.52	0.81	0.60	0.52	0.55	0.21	0.12	0.30	0.43	0.40	0.74	0.42	0.32	0.37	0.22
PixelOv	0.26	0.26	0.22	0.24	0.64	0.40	0.16	0.73	0.29	0.38	0.76	0.48	0.48	0.54	0.19	0.06	0.29	0.41	0.28	0.70	0.39	0.32	0.35	0.15
CTM	0.12	0.30	0.22	0.20	0.16	0.17	0.17	0.12	0.17	0.08	0.10	0.16	0.26	0.21	0.20	0.08	0.16	0.15	0.13	0.18	0.17	0.30	0.21	0.17
CTD	0.66	0.47	0.41	0.48	1.00	0.88	0.47	1.00	0.60	1.00	1.00	0.87	0.69	0.86	0.39	0.03	0.55	0.95	0.84	0.97	0.68	0.50	0.76	0.35
TC	0.46	0.56	0.42	0.37	0.65	0.56	0.47	0.78	0.47	0.51	0.81	0.60	0.67	0.56	0.37	0.10	0.30	0.42	0.50	0.76	0.48	0.53	0.43	0.40
CoTPSInv																								

Table B.2: Results for MS tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.16	0.31	0.37	0.20	0.15	0.38	0.18	0.44	0.36	0.26	0.55	0.45	0.47	0.30	0.28	0.11	0.29	0.51	0.21	0.33	0.47	0.25	0.37	0.25
ATA	0.16	0.31	0.37	0.20	0.15	0.39	0.19	0.44	0.36	0.26	0.56	0.46	0.47	0.30	0.28	0.11	0.29	0.51	0.22	0.34	0.47	0.25	0.38	0.25
ATEinv	0.75	0.46	0.50	0.65	0.46	0.54	0.57	0.95	0.48	0.98	0.96	0.59	0.84	0.72	0.38	0.99	0.57	0.61	0.81	0.47	0.62	0.28	0.46	0.38
AUCinv	0.24	0.40	0.48	0.20	0.16	0.56	0.21	0.45	0.47	0.28	0.57	0.58	0.50	0.31	0.33	0.12	0.30	0.53	0.23	0.34	0.50	0.25	0.41	0.38
PixelOv	0.24	0.40	0.49	0.20	0.16	0.56	0.21	0.46	0.48	0.28	0.58	0.59	0.50	0.31	0.33	0.13	0.30	0.54	0.24	0.34	0.50	0.25	0.42	0.38
CTM	0.16	0.31	0.37	0.20	0.15	0.39	0.19	0.44	0.36	0.26	0.56	0.46	0.47	0.30	0.28	0.11	0.29	0.51	0.22	0.34	0.47	0.25	0.38	0.25
CTD	0.17	0.23	0.16	0.27	0.23	0.18	0.23	0.17	0.19	0.18	0.11	0.14	0.26	0.21	0.19	0.14	0.18	0.13	0.21	0.19	0.17	0.33	0.20	0.15
TC	0.24	0.63	0.81	0.33	0.26	0.88	0.35	0.94	0.78	0.48	1.00	0.97	0.86	0.66	0.58	0.16	0.55	0.96	0.36	0.53	0.81	0.37	0.70	0.58
CoTPSinv	0.23	0.49	0.51	0.32	0.29	0.56	0.40	0.45	0.55	0.27	0.57	0.58	0.55	0.39	0.46	0.11	0.33	0.54	0.35	0.50	0.51	0.48	0.49	0.48

Table B.3: Results for PFC tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.32	0.10	0.22	0.45	0.03	0.36	0.38	0.18	0.33	0.38	0.23	0.38	0.37	0.12	0.20	0.15	0.18	0.28	0.37	0.20	0.35	0.17	0.07	0.09
ATA	0.33	0.10	0.23	0.46	0.03	0.36	0.38	0.18	0.34	0.39	0.23	0.39	0.37	0.12	0.20	0.15	0.18	0.28	0.38	0.20	0.35	0.17	0.07	0.09
ATEinv	0.72	0.21	0.49	0.94	0.07	0.77	0.83	0.35	0.72	0.83	0.45	0.83	0.78	0.24	0.43	0.91	0.33	0.45	0.82	0.41	0.72	0.35	0.15	0.18
AUCinv	0.47	0.11	0.26	0.47	0.03	0.41	0.48	0.18	0.36	0.48	0.24	0.42	0.35	0.12	0.22	0.18	0.18	0.28	0.47	0.20	0.37	0.16	0.07	0.10
PixelOv	0.48	0.11	0.27	0.48	0.04	0.42	0.48	0.18	0.36	0.49	0.24	0.43	0.35	0.12	0.22	0.18	0.18	0.28	0.47	0.20	0.37	0.17	0.07	0.11
CTM	0.33	0.10	0.23	0.46	0.03	0.36	0.38	0.18	0.34	0.39	0.23	0.39	0.37	0.12	0.20	0.15	0.18	0.28	0.38	0.20	0.35	0.17	0.07	0.09
CTD	0.08	0.15	0.09	0.02	0.12	0.08	0.02	0.21	0.10	0.02	0.22	0.07	0.14	0.17	0.13	0.11	0.13	0.18	0.02	0.19	0.09	0.21	0.14	0.13
TC	0.81	0.22	0.61	1.00	0.07	0.95	1.00	0.40	0.85	1.00	0.52	0.98	0.82	0.26	0.51	0.46	0.39	0.57	1.00	0.43	0.84	0.37	0.17	0.23
CoTPSinv	0.47	0.26	0.31	0.47	0.11	0.41	0.47	0.41	0.37	0.48	0.48	0.42	0.43	0.30	0.34	0.17	0.34	0.40	0.47	0.38	0.41	0.41	0.20	0.26

Table B.4: Results for LK tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.43	0.01	0.17	0.88	0.45	0.48	0.66	0.62	0.32	0.72	0.63	0.52	0.53	0.21	0.23	0.46	0.26	0.40	0.71	0.70	0.42	0.35	0.02	0.01
ATA	0.44	0.01	0.17	0.89	0.45	0.48	0.66	0.62	0.33	0.73	0.63	0.53	0.54	0.21	0.23	0.46	0.26	0.40	0.71	0.70	0.42	0.35	0.02	0.01
ATEinv	0.51	0.23	0.18	0.93	0.97	0.66	0.76	0.96	0.46	0.83	0.97	0.75	0.63	0.76	0.44	0.58	0.52	0.63	0.84	0.94	0.65	0.96	0.42	0.16
AUCinv	0.62	0.01	0.19	0.92	0.47	0.63	0.86	0.64	0.42	0.92	0.65	0.63	0.55	0.20	0.25	0.51	0.27	0.41	0.93	0.71	0.43	0.35	0.02	0.02
PixelOv	0.62	0.01	0.19	0.93	0.47	0.64	0.87	0.65	0.42	0.93	0.65	0.63	0.56	0.21	0.25	0.51	0.27	0.41	0.94	0.72	0.44	0.35	0.03	0.02
CTM	0.44	0.01	0.17	0.89	0.45	0.48	0.66	0.62	0.33	0.73	0.63	0.53	0.54	0.21	0.23	0.46	0.26	0.40	0.71	0.70	0.42	0.35	0.02	0.01
CTD	0.11	0.01	0.15	0.05	0.05	0.17	0.06	0.07	0.12	0.03	0.07	0.11	0.21	0.07	0.18	0.22	0.18	0.12	0.03	0.08	0.10	0.06	0.02	0.02
TC	0.63	0.00	0.49	1.00	1.00	0.94	1.00	1.00	0.73	1.00	1.00	0.98	0.74	0.41	0.45	0.76	0.55	0.75	1.00	0.98	0.78	0.67	0.01	0.01
CoTPSinv	0.71	0.19	0.34	0.92	0.47	0.63	0.86	0.64	0.46	0.92	0.65	0.62	0.69	0.30	0.39	0.50	0.31	0.42	0.93	0.71	0.43	0.35	0.15	0.17

Table B.5: Results for IVT tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.54	0.27	0.33	0.78	0.71	0.33	0.69	0.70	0.32	0.67	0.74	0.28	0.43	0.44	0.19	0.00	0.24	0.38	0.62	0.73	0.26	0.34	0.27	0.05
ATA	0.54	0.27	0.34	0.78	0.71	0.33	0.70	0.70	0.32	0.67	0.75	0.29	0.44	0.44	0.19	0.00	0.24	0.38	0.63	0.73	0.26	0.34	0.27	0.05
ATEinv	0.67	0.63	0.37	0.80	0.86	0.41	0.79	0.91	0.40	0.78	0.87	0.48	0.71	0.74	0.62	1.00	0.46	0.70	0.72	0.92	0.54	0.71	0.51	0.65
AUCinv	0.89	0.30	0.39	0.81	0.75	0.46	0.83	0.74	0.41	0.79	0.80	0.35	0.43	0.44	0.21	0.00	0.24	0.39	0.82	0.74	0.27	0.33	0.28	0.06
PixelOv	0.90	0.30	0.39	0.82	0.76	0.46	0.83	0.74	0.41	0.80	0.80	0.35	0.44	0.44	0.21	0.00	0.25	0.39	0.83	0.74	0.27	0.33	0.29	0.06
CTM	0.54	0.27	0.34	0.78	0.71	0.33	0.70	0.70	0.32	0.67	0.75	0.29	0.44	0.44	0.19	0.00	0.24	0.38	0.63	0.73	0.26	0.34	0.27	0.05
CTD	0.14	0.14	0.22	0.04	0.12	0.17	0.05	0.09	0.17	0.04	0.10	0.11	0.20	0.12	0.15	0.00	0.16	0.17	0.05	0.10	0.19	0.22	0.22	0.06
TC	0.87	0.47	0.63	1.00	1.00	0.71	1.00	1.00	0.66	1.00	1.00	0.54	0.73	0.97	0.36	0.00	0.44	0.72	1.00	1.00	0.46	0.58	0.47	0.16
CoTPSinv	0.89	0.38	0.55	0.81	0.75	0.57	0.82	0.74	0.52	0.79	0.79	0.47	0.61	0.45	0.33	0.00	0.24	0.49	0.82	0.74	0.37	0.47	0.45	0.15

Table B.6: Results for TLD tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.34	0.38	0.47	0.27	0.70	0.54	0.19	0.74	0.43	0.63	0.76	0.59	0.57	0.52	0.27	0.05	0.28	0.41	0.35	0.37	0.46	0.41	0.35	0.21
ATA	0.35	0.38	0.48	0.27	0.70	0.54	0.19	0.74	0.44	0.64	0.76	0.60	0.57	0.52	0.28	0.05	0.28	0.41	0.35	0.37	0.46	0.41	0.36	0.21
ATEinv	0.48	0.46	0.63	0.35	0.77	0.67	0.27	0.80	0.55	0.78	0.81	0.74	0.65	0.59	0.35	0.19	0.34	0.44	0.50	0.45	0.56	0.47	0.45	0.34
AUCinv	0.54	0.50	0.56	0.28	0.75	0.70	0.27	0.78	0.52	0.84	0.80	0.69	0.60	0.52	0.30	0.10	0.28	0.41	0.48	0.36	0.49	0.41	0.39	0.28
PixelOv	0.54	0.50	0.57	0.28	0.75	0.70	0.27	0.78	0.53	0.85	0.80	0.70	0.61	0.53	0.30	0.11	0.29	0.42	0.48	0.36	0.49	0.41	0.40	0.29
CTM	0.35	0.38	0.48	0.27	0.70	0.54	0.19	0.74	0.44	0.64	0.76	0.60	0.57	0.52	0.28	0.05	0.28	0.41	0.35	0.37	0.46	0.41	0.36	0.21
CTD	0.14	0.30	0.14	0.20	0.15	0.19	0.21	0.11	0.16	0.03	0.10	0.14	0.24	0.21	0.23	0.08	0.16	0.20	0.14	0.34	0.17	0.25	0.22	0.14
TC	0.76	0.58	0.96	0.50	1.00	0.96	0.47	1.00	0.77	1.00	1.00	0.97	0.77	0.84	0.50	0.03	0.56	0.84	0.85	0.52	0.75	0.72	0.69	0.51
CoTPSinv	0.58	0.66	0.55	0.40	0.75	0.69	0.52	0.78	0.58	0.84	0.80	0.69	0.70	0.60	0.45	0.09	0.28	0.46	0.58	0.53	0.52	0.51	0.49	0.39

Table B.7: Results for CBWH tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.44	0.04	0.46	0.33	0.03	0.28	0.31	0.39	0.32	0.53	0.60	0.42	0.37	0.26	0.19	0.52	0.28	0.41	0.64	0.67	0.41	0.20	0.12	0.15
ATA	0.45	0.04	0.46	0.33	0.03	0.28	0.31	0.39	0.33	0.53	0.60	0.42	0.37	0.26	0.19	0.53	0.28	0.41	0.65	0.67	0.41	0.20	0.13	0.15
ATEinv	0.47	0.79	0.70	1.00	0.99	0.66	0.85	0.96	0.57	0.90	0.97	0.71	0.62	0.89	0.47	0.58	0.70	0.60	0.84	0.82	0.69	0.46	0.42	0.44
AUCinv	0.51	0.04	0.56	0.33	0.03	0.36	0.38	0.40	0.39	0.62	0.62	0.52	0.38	0.26	0.20	0.53	0.29	0.42	0.79	0.68	0.42	0.20	0.14	0.19
PixelOv	0.52	0.04	0.56	0.33	0.03	0.37	0.38	0.41	0.39	0.62	0.62	0.52	0.38	0.26	0.20	0.54	0.29	0.42	0.79	0.69	0.42	0.20	0.14	0.20
CTM	0.45	0.04	0.46	0.33	0.03	0.28	0.31	0.39	0.33	0.53	0.60	0.42	0.37	0.26	0.19	0.53	0.28	0.41	0.65	0.67	0.41	0.20	0.13	0.15
CTD	0.06	0.10	0.11	0.23	0.11	0.20	0.33	0.21	0.16	0.03	0.10	0.11	0.25	0.18	0.18	0.08	0.19	0.10	0.09	0.13	0.15	0.28	0.18	0.13
TC	0.95	0.06	0.98	0.43	0.05	0.64	0.47	0.74	0.67	1.00	1.00	0.88	0.63	0.62	0.43	1.00	0.52	0.84	0.96	0.98	0.88	0.34	0.25	0.34
CoTPSinv	0.51	0.17	0.55	0.44	0.15	0.48	0.63	0.40	0.42	0.62	0.62	0.52	0.58	0.33	0.38	0.53	0.34	0.42	0.81	0.68	0.44	0.40	0.29	0.30

Table B.8: Results for SOAMST tracker

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.40	0.08	0.23	0.47	0.31	0.27	0.28	0.50	0.14	0.57	0.55	0.37	0.54	0.38	0.25	0.31	0.35	0.43	0.45	0.52	0.41	0.25	0.22	0.11
ATA	0.41	0.08	0.23	0.47	0.31	0.28	0.28	0.50	0.14	0.57	0.56	0.38	0.55	0.38	0.25	0.31	0.35	0.43	0.46	0.52	0.42	0.25	0.22	0.11
ATEinv	0.55	0.54	0.28	0.67	0.49	0.42	0.70	0.74	0.18	0.83	0.75	0.53	0.83	0.62	0.33	0.73	0.46	0.51	0.67	0.69	0.54	0.35	0.36	0.19
AUCinv	0.58	0.09	0.28	0.51	0.31	0.39	0.40	0.52	0.18	0.79	0.58	0.47	0.56	0.38	0.28	0.41	0.35	0.44	0.64	0.53	0.44	0.25	0.24	0.13
PixelOv	0.59	0.09	0.29	0.51	0.31	0.40	0.40	0.53	0.18	0.79	0.58	0.48	0.57	0.39	0.28	0.42	0.36	0.44	0.64	0.53	0.45	0.25	0.25	0.14
CTM	0.41	0.08	0.23	0.47	0.31	0.28	0.28	0.50	0.14	0.57	0.56	0.38	0.55	0.38	0.25	0.31	0.35	0.43	0.46	0.52	0.42	0.25	0.22	0.11
CTD	0.12	0.20	0.25	0.18	0.19	0.24	0.28	0.20	0.20	0.03	0.21	0.20	0.25	0.18	0.23	0.11	0.17	0.19	0.15	0.25	0.20	0.31	0.22	0.14
TC	0.71	0.15	0.41	0.80	0.67	0.55	0.51	0.92	0.24	1.00	0.96	0.70	0.73	0.81	0.45	0.82	0.67	0.77	0.87	0.83	0.68	0.38	0.42	0.24
CoTPSinv	0.65	0.25	0.46	0.53	0.31	0.52	0.65	0.52	0.41	0.78	0.58	0.55	0.67	0.42	0.45	0.41	0.35	0.48	0.68	0.59	0.46	0.49	0.39	0.29

Table B.9: Results for  $\geq 1$  (OR) fusion



	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.54	0.12	0.40	0.88	0.70	0.45	0.42	0.72	0.41	0.69	0.76	0.57	0.65	0.51	0.29	0.31	0.35	0.50	0.67	0.71	0.55	0.38	0.39	0.19
ATA	0.54	0.12	0.41	0.89	0.70	0.46	0.43	0.72	0.41	0.70	0.76	0.58	0.66	0.51	0.30	0.31	0.35	0.50	0.68	0.71	0.56	0.38	0.40	0.19
ATEinv	0.65	0.57	0.52	0.94	0.81	0.62	0.83	0.89	0.52	0.83	0.88	0.72	0.89	0.72	0.37	0.95	0.42	0.56	0.79	0.81	0.68	0.45	0.52	0.28
AUCinv	0.89	0.14	0.53	0.93	0.75	0.64	0.57	0.76	0.54	0.94	0.81	0.73	0.70	0.52	0.33	0.37	0.35	0.52	0.90	0.73	0.59	0.39	0.43	0.26
PixelOv	0.89	0.14	0.53	0.93	0.75	0.65	0.57	0.77	0.55	0.94	0.81	0.74	0.70	0.52	0.33	0.38	0.36	0.52	0.91	0.73	0.60	0.39	0.44	0.26
CTM	0.54	0.12	0.41	0.89	0.70	0.46	0.43	0.72	0.41	0.70	0.76	0.58	0.66	0.51	0.30	0.31	0.35	0.50	0.68	0.71	0.56	0.38	0.40	0.19
CTD	0.15	0.26	0.23	0.07	0.13	0.19	0.29	0.10	0.20	0.04	0.10	0.14	0.23	0.19	0.24	0.20	0.16	0.17	0.09	0.20	0.17	0.34	0.23	0.17
TC	0.84	0.17	0.77	1.00	1.00	0.90	0.60	1.00	0.76	1.00	1.00	0.98	0.81	0.85	0.51	0.56	0.63	0.94	0.96	0.92	0.90	0.55	0.70	0.41
CoTPSinv	0.92	0.31	0.64	0.93	0.75	0.67	0.77	0.76	0.61	0.94	0.81	0.73	0.80	0.57	0.48	0.37	0.35	0.54	0.92	0.77	0.59	0.60	0.52	0.41

Table B.10: Results for  $\geq 2$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.35	0.10	0.16	0.40	0.16	0.22	0.33	0.40	0.10	0.58	0.44	0.30	0.43	0.29	0.21	0.39	0.23	0.25	0.47	0.39	0.31	0.23	0.14	0.09
ATA	0.35	0.10	0.16	0.41	0.16	0.22	0.33	0.40	0.10	0.58	0.45	0.31	0.44	0.29	0.21	0.39	0.23	0.25	0.47	0.39	0.32	0.23	0.14	0.10
ATEinv	0.35	0.10	0.17	0.41	0.16	0.25	0.35	0.41	0.11	0.62	0.46	0.32	0.44	0.29	0.22	0.41	0.24	0.26	0.50	0.39	0.32	0.24	0.15	0.10
AUCinv	0.37	0.11	0.17	0.40	0.16	0.28	0.37	0.41	0.11	0.67	0.45	0.34	0.43	0.29	0.22	0.38	0.23	0.25	0.54	0.38	0.32	0.23	0.14	0.11
PixelOv	0.37	0.11	0.18	0.40	0.16	0.28	0.37	0.41	0.12	0.68	0.46	0.34	0.44	0.29	0.22	0.39	0.23	0.26	0.55	0.39	0.32	0.23	0.15	0.11
CTM	0.35	0.10	0.16	0.41	0.16	0.22	0.33	0.40	0.10	0.58	0.45	0.31	0.44	0.29	0.21	0.39	0.23	0.25	0.47	0.39	0.32	0.23	0.14	0.10
CTD	0.07	0.15	0.12	0.17	0.20	0.20	0.23	0.29	0.11	0.03	0.31	0.15	0.18	0.23	0.15	0.18	0.14	0.14	0.15	0.24	0.13	0.28	0.15	0.10
TC	0.69	0.13	0.34	0.64	0.22	0.40	0.53	0.55	0.14	1.00	0.55	0.64	0.73	0.46	0.43	0.76	0.41	0.55	0.84	0.65	0.59	0.37	0.15	0.16
CoTPSinv	0.36	0.11	0.16	0.39	0.15	0.27	0.36	0.41	0.13	0.67	0.45	0.33	0.43	0.28	0.29	0.38	0.23	0.25	0.54	0.38	0.31	0.23	0.14	0.16

Table B.11: Results for  $\geq 3$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.44	0.31	0.29	0.56	0.69	0.43	0.35	0.74	0.40	0.69	0.75	0.50	0.62	0.55	0.27	0.54	0.28	0.36	0.52	0.60	0.42	0.31	0.29	0.17
ATA	0.44	0.31	0.30	0.56	0.69	0.44	0.36	0.75	0.40	0.70	0.75	0.51	0.63	0.55	0.27	0.54	0.28	0.36	0.53	0.60	0.43	0.31	0.30	0.18
ATEinv	0.48	0.36	0.31	0.57	0.73	0.51	0.38	0.78	0.44	0.76	0.78	0.56	0.64	0.60	0.30	0.63	0.32	0.37	0.57	0.62	0.45	0.32	0.31	0.20
AUCinv	0.61	0.37	0.33	0.56	0.72	0.57	0.41	0.78	0.49	0.84	0.78	0.60	0.62	0.56	0.29	0.58	0.28	0.37	0.64	0.60	0.44	0.31	0.31	0.23
PixelOv	0.62	0.38	0.33	0.57	0.72	0.58	0.42	0.78	0.49	0.84	0.79	0.60	0.63	0.56	0.30	0.58	0.29	0.37	0.65	0.61	0.45	0.31	0.32	0.23
CTM	0.44	0.31	0.30	0.56	0.69	0.44	0.36	0.75	0.40	0.70	0.75	0.51	0.63	0.55	0.27	0.54	0.28	0.36	0.53	0.60	0.43	0.31	0.30	0.18
CTD	0.11	0.26	0.16	0.16	0.11	0.16	0.24	0.09	0.16	0.02	0.09	0.11	0.17	0.20	0.20	0.10	0.15	0.18	0.13	0.20	0.16	0.30	0.18	0.14
TC	0.82	0.52	0.65	0.90	1.00	0.87	0.55	1.00	0.77	1.00	1.00	0.98	0.97	0.84	0.49	1.00	0.50	0.67	0.87	0.97	0.75	0.43	0.57	0.36
CoTPSinv	0.61	0.51	0.32	0.56	0.72	0.57	0.41	0.78	0.51	0.84	0.78	0.59	0.62	0.56	0.39	0.57	0.28	0.36	0.64	0.60	0.44	0.36	0.31	0.34

Table B.12: Results for  $\geq 4$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.46	0.31	0.45	0.84	0.72	0.47	0.67	0.78	0.44	0.70	0.79	0.56	0.66	0.54	0.27	0.32	0.31	0.46	0.63	0.71	0.52	0.41	0.39	0.20
ATA	0.47	0.31	0.46	0.84	0.72	0.48	0.67	0.78	0.44	0.71	0.79	0.57	0.67	0.54	0.28	0.32	0.31	0.46	0.63	0.72	0.53	0.41	0.39	0.21
ATEinv	0.53	0.58	0.53	0.88	0.83	0.60	0.77	0.86	0.52	0.80	0.86	0.66	0.72	0.65	0.37	0.63	0.36	0.49	0.72	0.78	0.59	0.53	0.48	0.39
AUCinv	0.72	0.39	0.55	0.85	0.76	0.65	0.86	0.83	0.56	0.90	0.84	0.69	0.67	0.55	0.31	0.40	0.31	0.47	0.82	0.73	0.56	0.41	0.43	0.28
PixelOv	0.73	0.39	0.56	0.86	0.77	0.66	0.86	0.83	0.56	0.90	0.85	0.69	0.67	0.56	0.31	0.41	0.32	0.48	0.82	0.74	0.56	0.41	0.44	0.29
CTM	0.47	0.31	0.46	0.84	0.72	0.48	0.67	0.78	0.44	0.71	0.79	0.57	0.67	0.54	0.28	0.32	0.31	0.46	0.63	0.72	0.53	0.41	0.39	0.21
CTD	0.13	0.27	0.15	0.06	0.14	0.17	0.05	0.10	0.16	0.03	0.10	0.13	0.18	0.21	0.23	0.11	0.16	0.16	0.10	0.13	0.15	0.32	0.21	0.17
T C	0.83	0.51	0.93	1.00	1.00	0.95	1.00	1.00	0.81	1.00	1.00	0.98	0.89	0.85	0.49	0.90	0.56	0.85	0.95	1.00	0.99	0.66	0.72	0.47
CoTPSinv	0.72	0.54	0.54	0.85	0.76	0.65	0.85	0.83	0.58	0.89	0.84	0.68	0.73	0.58	0.46	0.40	0.31	0.47	0.82	0.73	0.55	0.57	0.50	0.43

Table B.13: Results for  $\geq 5$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.51	0.23	0.47	0.77	0.68	0.46	0.63	0.75	0.41	0.68	0.77	0.57	0.64	0.52	0.27	0.25	0.34	0.52	0.64	0.73	0.56	0.39	0.32	0.16
ATA	0.51	0.23	0.47	0.78	0.68	0.46	0.64	0.76	0.42	0.69	0.77	0.57	0.64	0.52	0.27	0.26	0.34	0.53	0.65	0.73	0.56	0.39	0.32	0.16
ATEinv	0.64	0.71	0.63	0.96	0.90	0.64	0.85	0.91	0.57	0.84	0.90	0.72	0.90	0.76	0.49	0.75	0.41	0.63	0.80	0.85	0.71	0.82	0.63	0.45
AUCinv	0.82	0.28	0.59	0.79	0.72	0.63	0.80	0.80	0.53	0.91	0.81	0.71	0.66	0.53	0.30	0.35	0.34	0.54	0.85	0.74	0.60	0.39	0.36	0.22
PixelOv	0.82	0.28	0.59	0.80	0.72	0.64	0.80	0.81	0.53	0.92	0.81	0.71	0.66	0.54	0.31	0.35	0.35	0.54	0.86	0.75	0.60	0.39	0.36	0.22
CTM	0.51	0.23	0.47	0.78	0.68	0.46	0.64	0.76	0.42	0.69	0.77	0.57	0.64	0.52	0.27	0.26	0.34	0.53	0.65	0.73	0.56	0.39	0.32	0.16
CTD	0.14	0.27	0.20	0.08	0.12	0.18	0.06	0.10	0.20	0.04	0.11	0.13	0.21	0.19	0.24	0.14	0.17	0.14	0.06	0.13	0.15	0.31	0.25	0.17
T C	0.82	0.40	0.86	1.00	1.00	0.93	1.00	1.00	0.78	1.00	1.00	0.98	0.86	0.84	0.46	0.56	0.60	0.92	1.00	0.99	0.97	0.64	0.55	0.33
CoTPSinv	0.82	0.45	0.62	0.79	0.71	0.63	0.80	0.80	0.59	0.91	0.81	0.70	0.74	0.58	0.47	0.34	0.34	0.54	0.85	0.74	0.59	0.59	0.51	0.40

Table B.14: Results for  $\geq 6$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.47	0.12	0.40	0.45	0.53	0.43	0.32	0.67	0.37	0.64	0.72	0.53	0.56	0.39	0.24	0.21	0.36	0.51	0.61	0.66	0.50	0.25	0.15	0.08
ATA	0.47	0.12	0.40	0.45	0.53	0.43	0.33	0.67	0.38	0.65	0.73	0.54	0.57	0.39	0.24	0.21	0.36	0.51	0.61	0.66	0.51	0.25	0.15	0.08
ATEinv	0.68	0.80	0.69	0.97	0.96	0.67	0.83	0.95	0.62	0.86	0.95	0.77	0.93	0.85	0.56	0.88	0.50	0.74	0.85	0.92	0.76	0.93	0.73	0.52
AUCinv	0.72	0.14	0.51	0.47	0.55	0.59	0.43	0.70	0.47	0.84	0.76	0.66	0.58	0.39	0.27	0.26	0.36	0.53	0.81	0.67	0.53	0.25	0.17	0.11
PixelOv	0.72	0.14	0.51	0.47	0.55	0.59	0.43	0.71	0.47	0.84	0.76	0.66	0.58	0.40	0.27	0.27	0.37	0.53	0.81	0.67	0.54	0.25	0.17	0.11
CTM	0.47	0.12	0.40	0.45	0.53	0.43	0.33	0.67	0.38	0.65	0.73	0.54	0.57	0.39	0.24	0.21	0.36	0.51	0.61	0.66	0.51	0.25	0.15	0.08
CTD	0.13	0.21	0.24	0.25	0.09	0.18	0.31	0.09	0.19	0.03	0.09	0.13	0.25	0.18	0.22	0.15	0.19	0.15	0.09	0.20	0.17	0.30	0.22	0.13
T C	0.77	0.20	0.77	0.62	1.00	0.92	0.50	1.00	0.74	1.00	1.00	0.98	0.75	0.74	0.44	0.50	0.68	0.92	0.96	0.90	0.91	0.39	0.27	0.21
CoTPSinv	0.76	0.34	0.64	0.60	0.54	0.58	0.63	0.70	0.54	0.83	0.76	0.66	0.70	0.48	0.43	0.35	0.36	0.55	0.84	0.73	0.53	0.49	0.40	0.29

Table B.15: Results for  $\geq 7$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.36	0.05	0.23	0.31	0.15	0.35	0.26	0.51	0.29	0.54	0.63	0.46	0.47	0.24	0.20	0.04	0.28	0.42	0.45	0.52	0.36	0.17	0.07	0.04
ATA	0.36	0.05	0.24	0.32	0.15	0.35	0.26	0.51	0.29	0.54	0.63	0.46	0.47	0.24	0.21	0.04	0.29	0.42	0.46	0.53	0.37	0.17	0.07	0.04
ATEinv	0.66	0.78	0.70	0.99	0.99	0.72	0.84	0.97	0.70	0.88	0.96	0.82	0.97	0.90	0.61	0.85	0.59	0.81	0.86	0.94	0.77	0.97	0.77	0.63
AUCinv	0.55	0.06	0.30	0.31	0.15	0.47	0.33	0.53	0.36	0.66	0.66	0.55	0.47	0.23	0.23	0.05	0.29	0.43	0.57	0.53	0.38	0.17	0.07	0.05
PixelOv	0.55	0.06	0.31	0.31	0.15	0.47	0.33	0.53	0.36	0.66	0.66	0.56	0.47	0.24	0.23	0.06	0.29	0.44	0.57	0.53	0.38	0.17	0.07	0.05
CTM	0.36	0.05	0.24	0.32	0.15	0.35	0.26	0.51	0.29	0.54	0.63	0.46	0.47	0.24	0.21	0.04	0.29	0.42	0.46	0.53	0.37	0.17	0.07	0.04
CTD	0.12	0.13	0.24	0.23	0.19	0.16	0.28	0.12	0.16	0.05	0.08	0.11	0.24	0.18	0.19	0.08	0.19	0.14	0.16	0.19	0.18	0.24	0.16	0.09
TC	0.68	0.09	0.46	0.50	0.31	0.84	0.47	0.97	0.66	1.00	1.00	0.94	0.70	0.52	0.42	0.03	0.60	0.78	0.86	0.82	0.74	0.33	0.12	0.11
CoTPSinv	0.60	0.17	0.51	0.43	0.29	0.47	0.58	0.52	0.43	0.66	0.66	0.56	0.62	0.33	0.39	0.21	0.34	0.49	0.64	0.62	0.47	0.40	0.20	0.20

Table B.16: Results for  $\geq 8$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.22	0.02	0.14	0.20	0.04	0.17	0.19	0.32	0.18	0.34	0.49	0.29	0.34	0.11	0.15	0.01	0.21	0.32	0.27	0.32	0.22	0.11	0.03	0.01
ATA	0.22	0.02	0.15	0.20	0.04	0.17	0.19	0.32	0.18	0.34	0.49	0.29	0.34	0.11	0.16	0.01	0.21	0.33	0.27	0.33	0.22	0.11	0.03	0.01
ATEinv	0.67	0.91	0.73	1.00	1.00	0.62	0.83	0.98	0.78	0.91	0.98	0.80	0.99	0.96	0.80	0.92	0.73	0.88	0.84	0.98	0.87	0.99	0.95	0.70
AUCinv	0.30	0.02	0.18	0.20	0.04	0.23	0.24	0.32	0.22	0.40	0.50	0.34	0.34	0.10	0.17	0.01	0.21	0.33	0.34	0.32	0.22	0.11	0.03	0.01
PixelOv	0.30	0.02	0.18	0.20	0.04	0.23	0.24	0.33	0.22	0.40	0.51	0.35	0.34	0.11	0.18	0.01	0.21	0.34	0.34	0.33	0.23	0.11	0.03	0.01
CTM	0.22	0.02	0.15	0.20	0.04	0.17	0.19	0.32	0.18	0.34	0.49	0.29	0.34	0.11	0.16	0.01	0.21	0.33	0.27	0.33	0.22	0.11	0.03	0.01
CTD	0.10	0.06	0.17	0.21	0.11	0.20	0.21	0.15	0.14	0.09	0.07	0.13	0.23	0.10	0.14	0.05	0.17	0.13	0.14	0.24	0.19	0.16	0.09	0.03
TC	0.59	0.04	0.33	0.44	0.08	0.42	0.46	0.76	0.43	1.00	1.00	0.73	0.57	0.21	0.38	0.02	0.46	0.60	0.84	0.53	0.47	0.21	0.08	0.01
CoTPSinv	0.37	0.11	0.36	0.32	0.16	0.46	0.49	0.37	0.37	0.39	0.50	0.42	0.48	0.24	0.33	0.11	0.35	0.42	0.44	0.49	0.37	0.33	0.14	0.10

Table B.17: Results for  $\geq 9$  fusion

	Complex mov.			Illum. global			Illum. local			Noise			Occlusion			Scale change			Similar Obj.			L4		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	C	F	P
SFDA	0.08	0.00	0.06	0.08	0.01	0.08	0.09	0.12	0.03	0.17	0.18	0.07	0.14	0.03	0.08	0.00	0.09	0.16	0.13	0.10	0.07	0.06	0.01	0.00
ATA	0.08	0.00	0.07	0.08	0.01	0.08	0.09	0.12	0.04	0.17	0.18	0.07	0.14	0.03	0.09	0.00	0.09	0.16	0.13	0.10	0.07	0.06	0.01	0.00
ATEinv	0.73	1.00	0.75	1.00	1.00	0.73	0.95	0.99	0.65	0.99	0.98	0.82	1.00	0.99	0.89	1.00	0.85	0.93	0.98	0.99	0.97	1.00	0.98	0.91
AUCinv	0.10	0.00	0.08	0.08	0.01	0.11	0.10	0.12	0.04	0.18	0.18	0.09	0.14	0.03	0.09	0.00	0.09	0.16	0.13	0.10	0.07	0.06	0.01	0.00
PixelOv	0.10	0.00	0.08	0.08	0.01	0.11	0.10	0.12	0.04	0.18	0.19	0.09	0.14	0.03	0.09	0.00	0.09	0.16	0.13	0.10	0.07	0.06	0.01	0.00
CTM	0.08	0.00	0.07	0.08	0.01	0.08	0.09	0.12	0.04	0.17	0.18	0.07	0.14	0.03	0.09	0.00	0.09	0.16	0.13	0.10	0.07	0.06	0.01	0.00
CTD	0.07	0.01	0.11	0.12	0.06	0.12	0.10	0.15	0.08	0.07	0.18	0.10	0.16	0.04	0.08	0.00	0.10	0.11	0.09	0.15	0.09	0.10	0.02	0.00
TC	0.11	0.00	0.19	0.15	0.04	0.22	0.19	0.35	0.10	0.38	0.52	0.13	0.40	0.03	0.26	0.00	0.23	0.43	0.23	0.25	0.19	0.20	0.00	0.00
CoTPSinv	0.15	0.07	0.27	0.22	0.08	0.33	0.35	0.34	0.16	0.17	0.43	0.31	0.36	0.20	0.21	0.00	0.21	0.31	0.29	0.28	0.24	0.28	0.09	0.03

Table B.18: Results for  $\geq 10$  fusion



## Appendix C

# Trackers and fusion results: Bar figures

This appendix presents the obtained fusion results with bar figures. Four figures (one for each dataset category) are presented for each one of the nine metrics.

### C.1 Sequence Frame Detection Accuracy (SFDA)

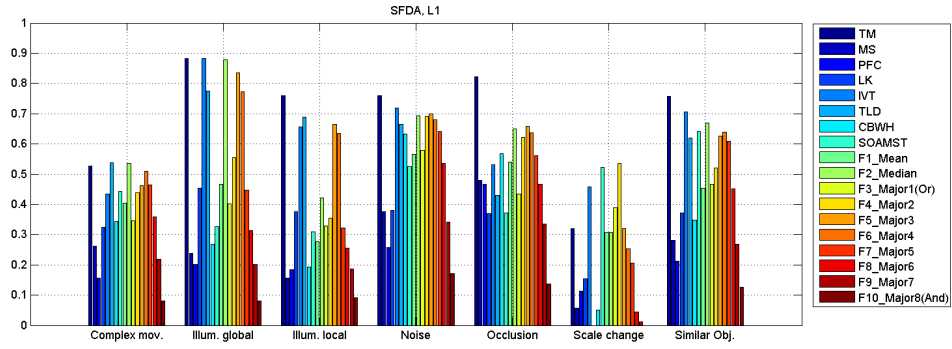


Figure C.1: Fusion SFDA result of L1

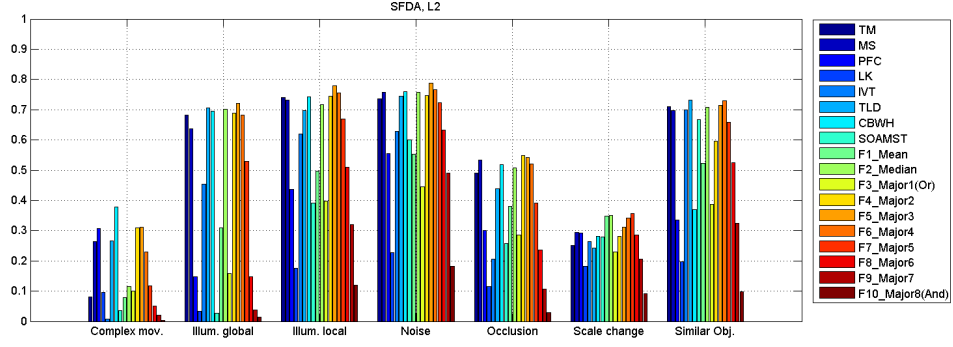


Figure C.2: Fusion SFDA result of L2

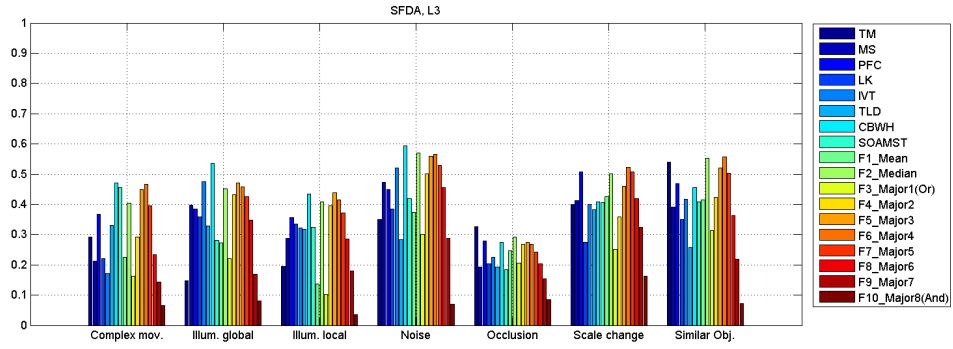


Figure C.3: Fusion SFDA result of L3

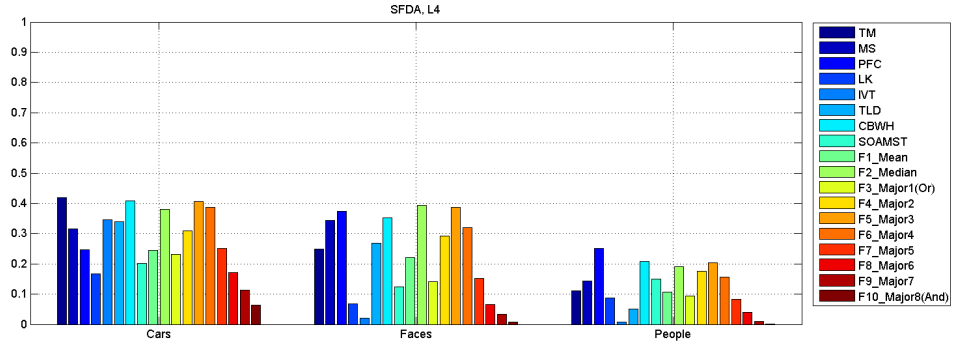


Figure C.4: Fusion SFDA result of L4

## C.2 Average Tracking Accuracy (ATA)

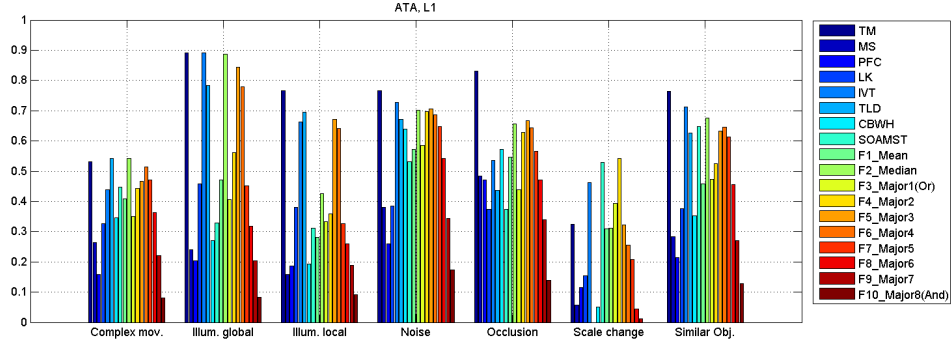


Figure C.5: Fusion ATA result of L1

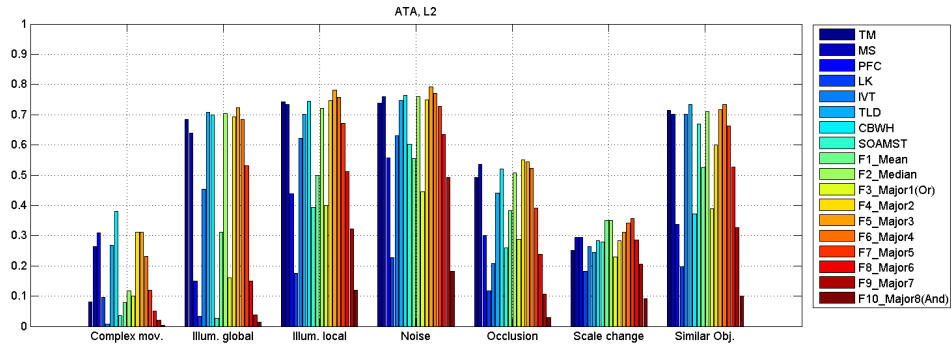


Figure C.6: Fusion ATA result of L2

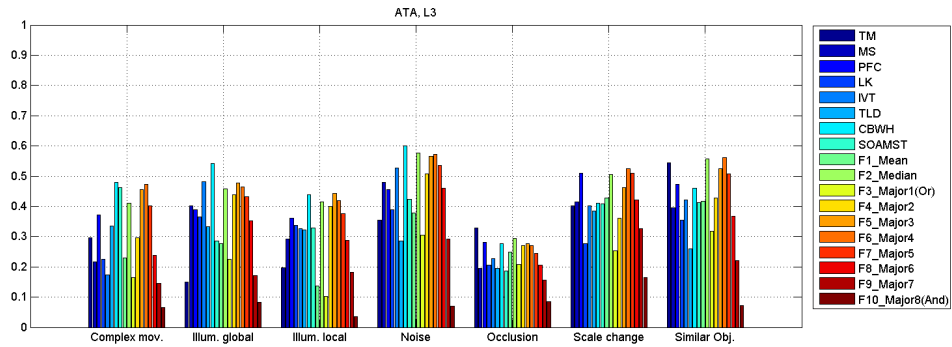


Figure C.7: Fusion ATA result of L3

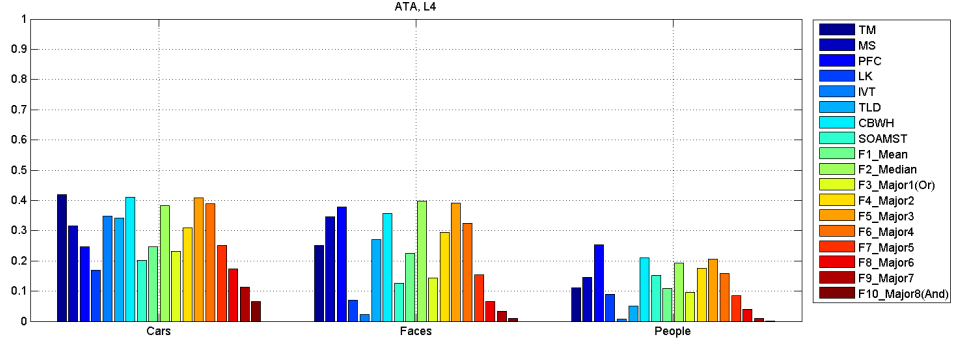


Figure C.8: Fusion ATA result of L4

### C.3 Average Tracking Error (ATE)

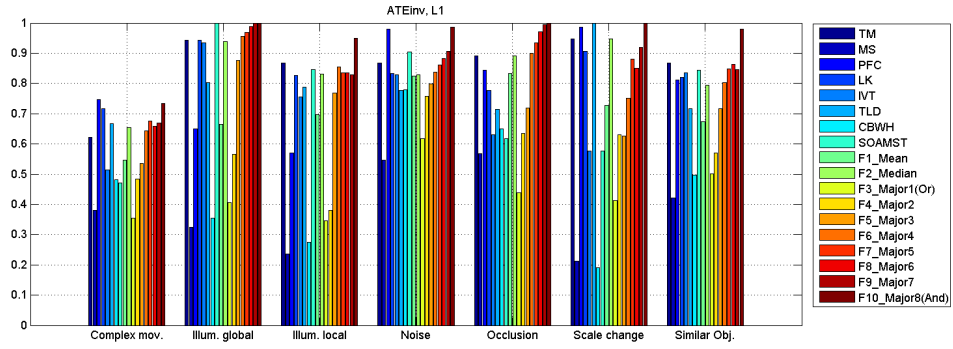


Figure C.9: Fusion ATEinv result of L1

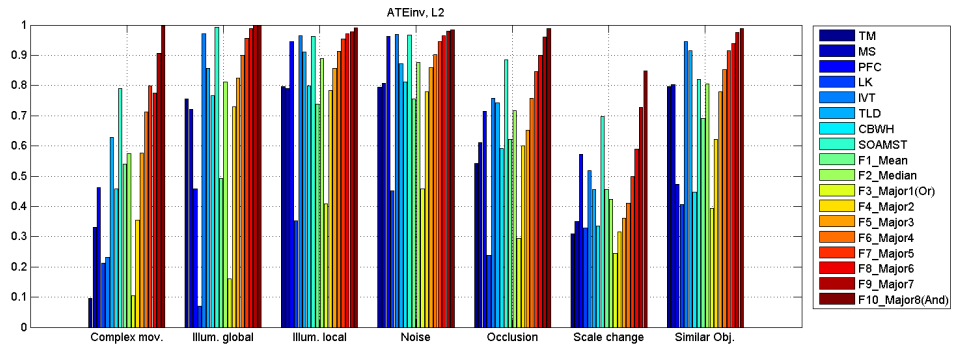


Figure C.10: Fusion ATEinv result of L2



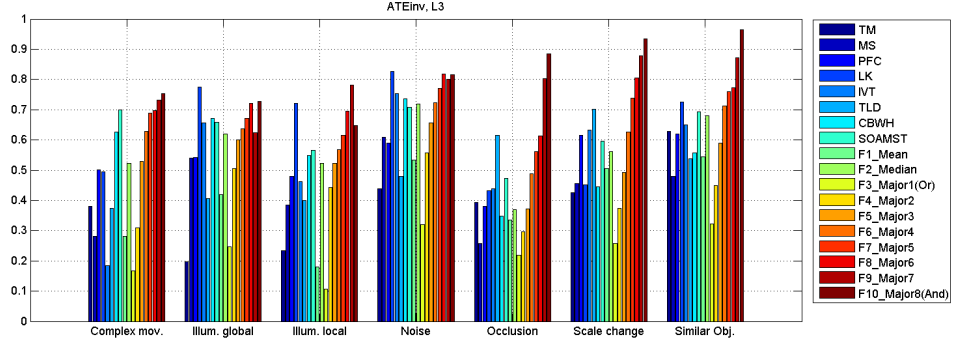


Figure C.11: Fusion ATEinv result of L3

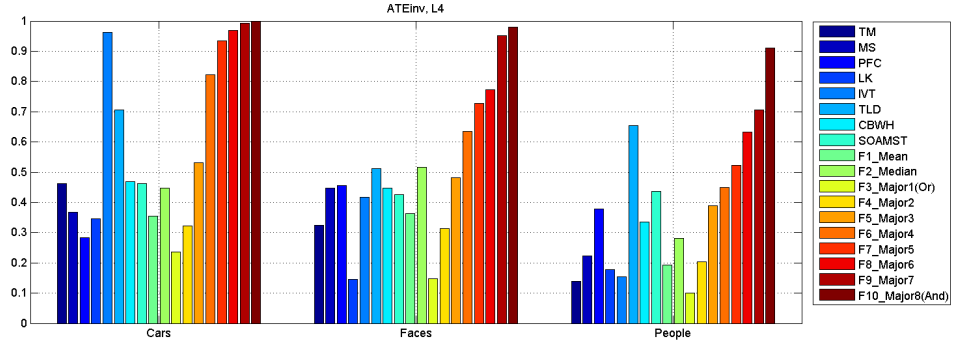


Figure C.12: Fusion ATEinv result of L4

## C.4 Overlap

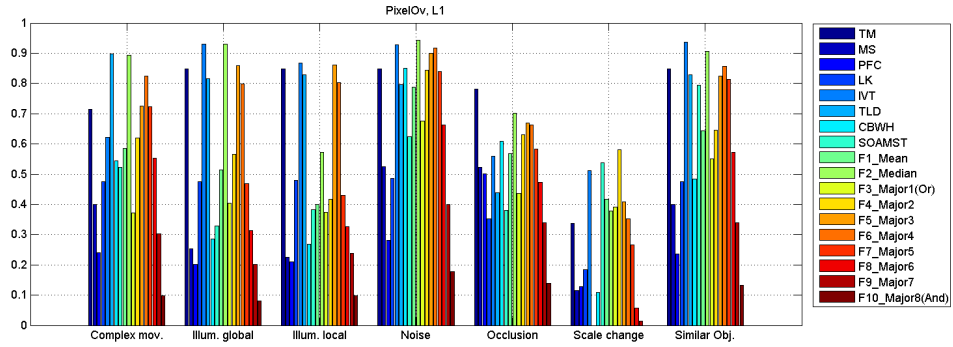


Figure C.13: Fusion PixelOverlap result of L1

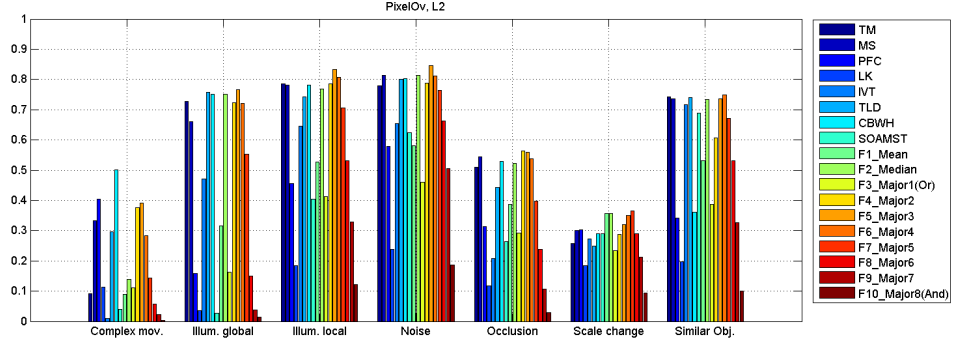


Figure C.14: Fusion PixelOverlap result of L2

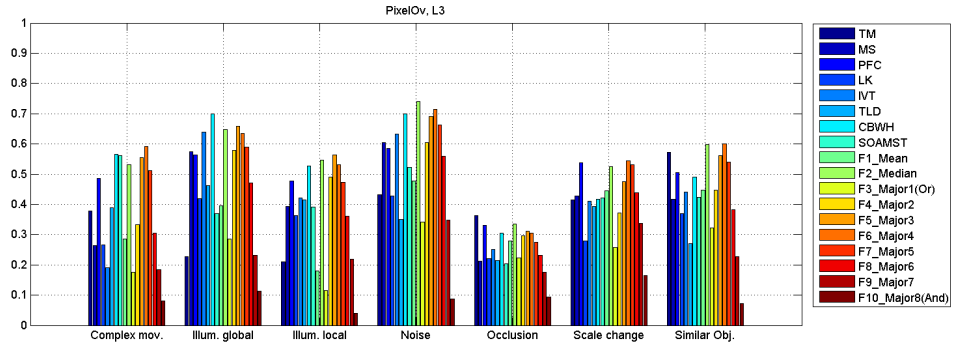


Figure C.15: Fusion PixelOverlap result of L3

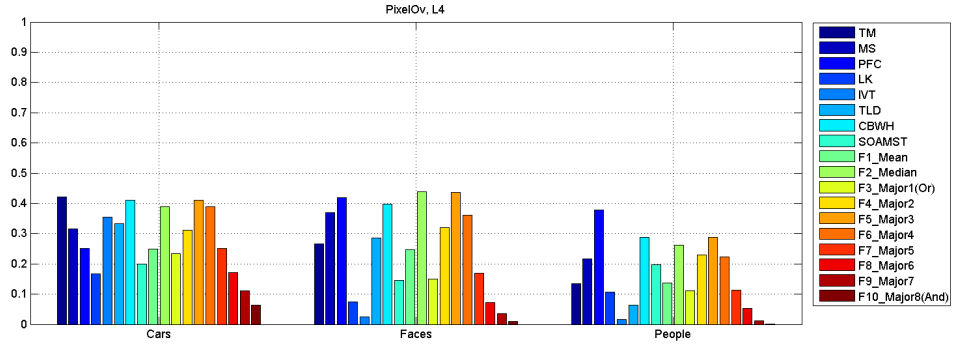


Figure C.16: Fusion PixelOverlap result of L4

## C.5 Area Under the lost track ratio Curve (AUC)

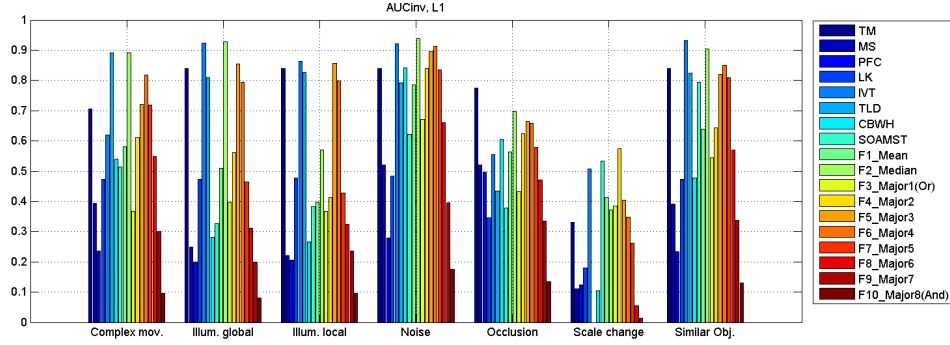


Figure C.17: Fusion AUCinv result of L1

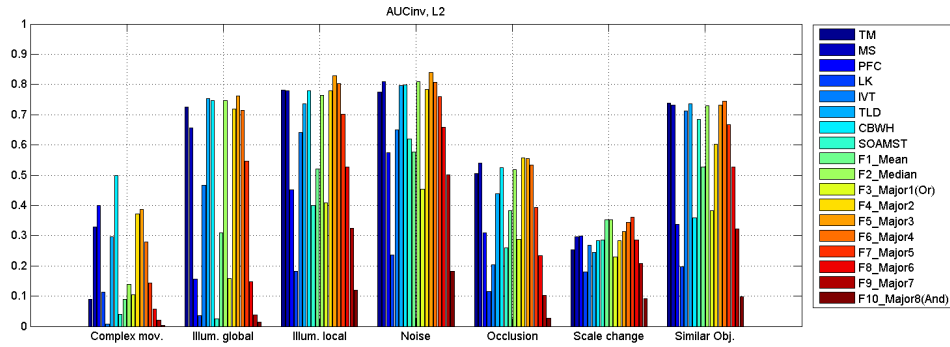


Figure C.18: Fusion AUCinv result of L2

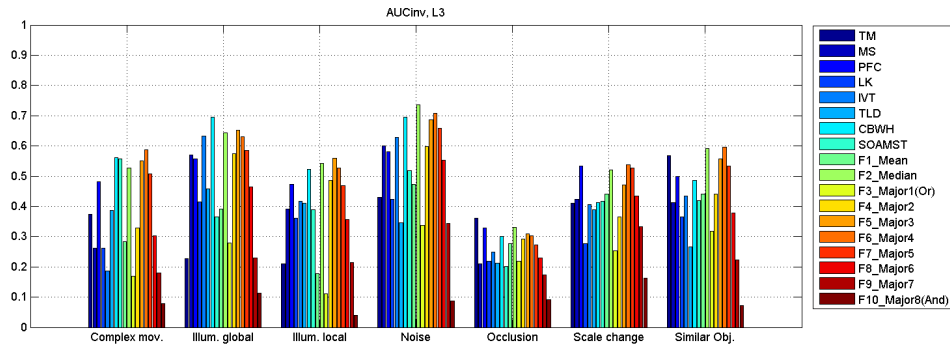


Figure C.19: Fusion AUCinv result of L3

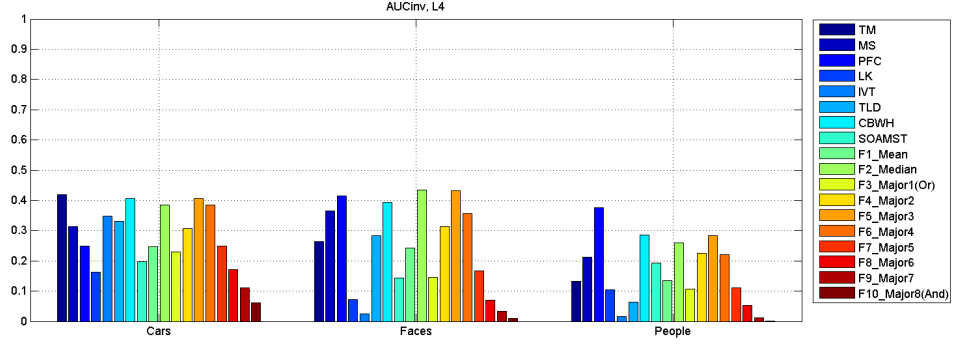


Figure C.20: Fusion AUCinv result of L4

## C.6 Closeness of Track (CT)

### C.6.1 The closeness of the whole sequence (CTM)

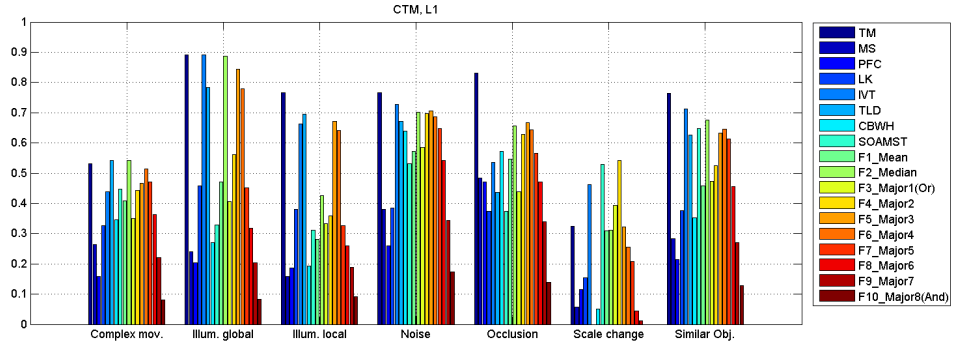


Figure C.21: Fusion CTM result of L1

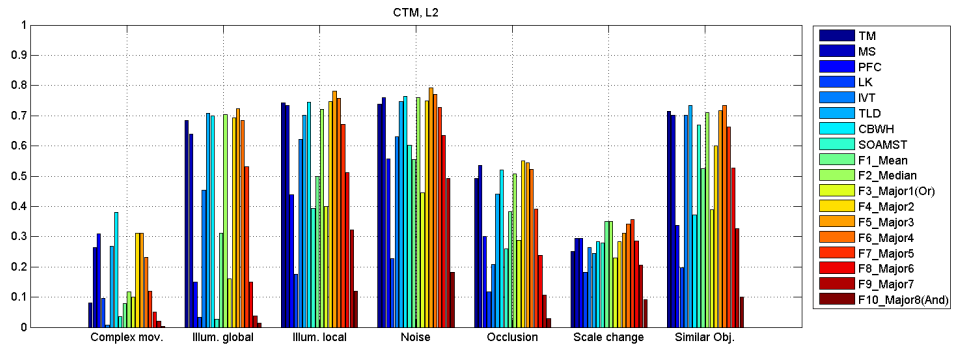


Figure C.22: Fusion CTM result of L2

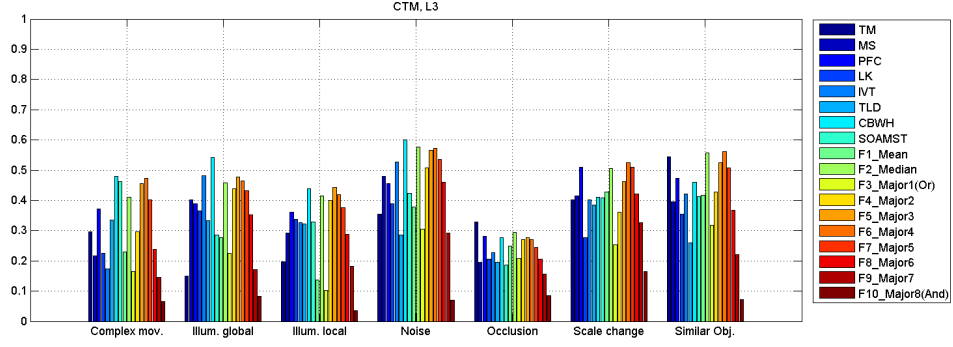


Figure C.23: Fusion CTM result of L3

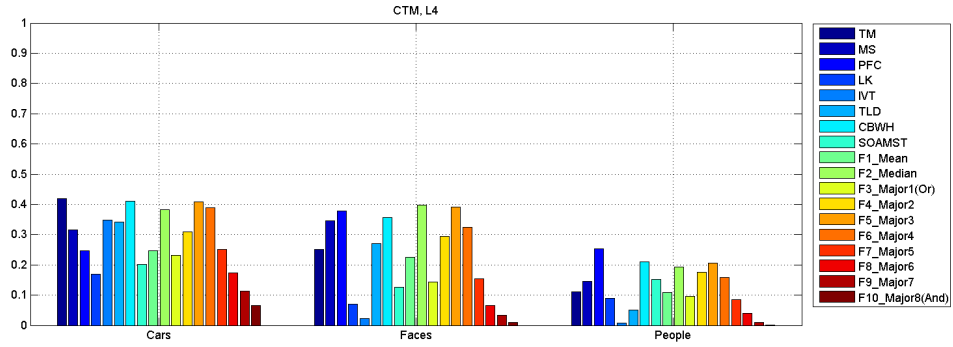


Figure C.24: Fusion CTM result of L4

### C.6.2 weighted standard deviation of track closeness (CTD)

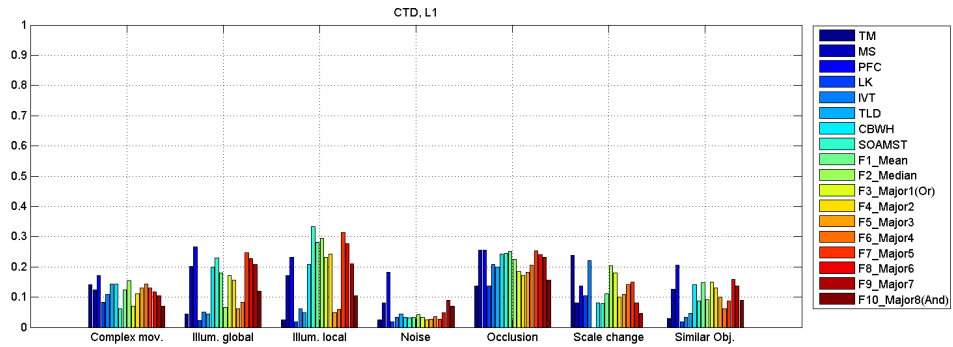


Figure C.25: Fusion CTD result of L1

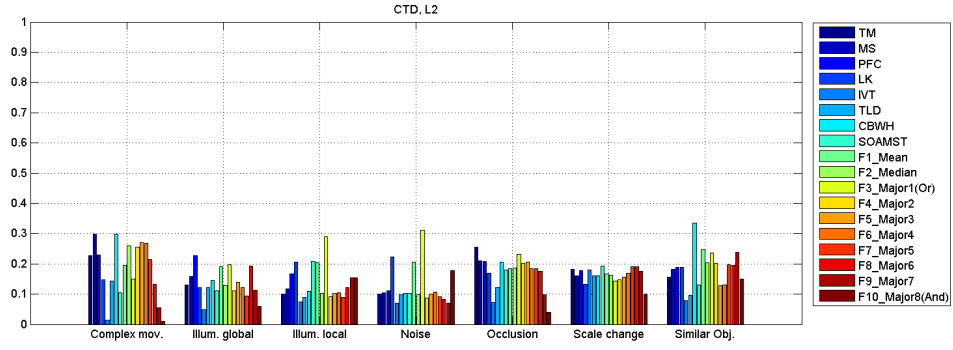


Figure C.26: Fusion CTD result of L2

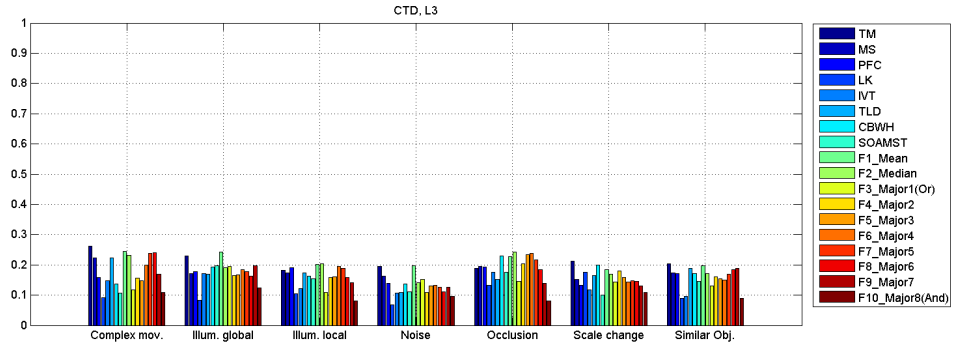


Figure C.27: Fusion CTD result of L3

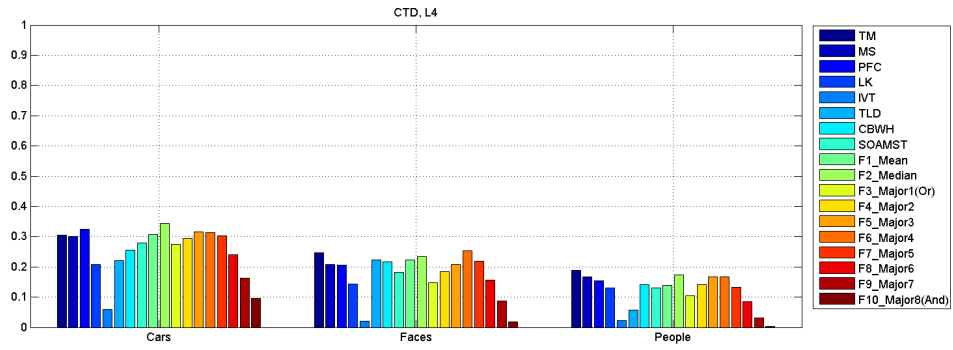


Figure C.28: Fusion CTD result of L4

## C.7 Track Completeness (TC)

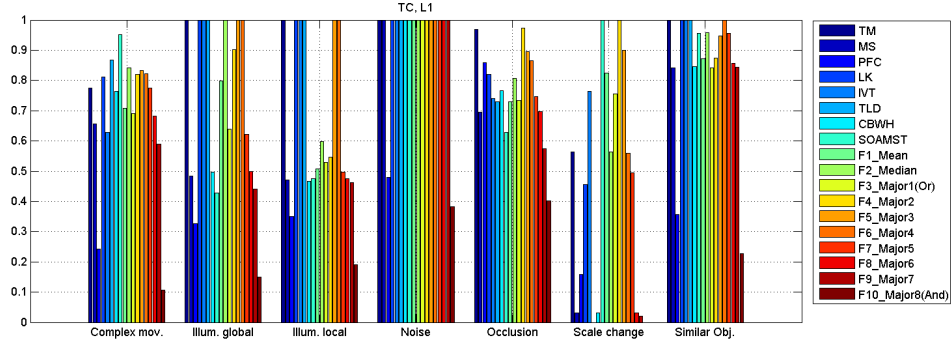


Figure C.29: Fusion TC result of L1

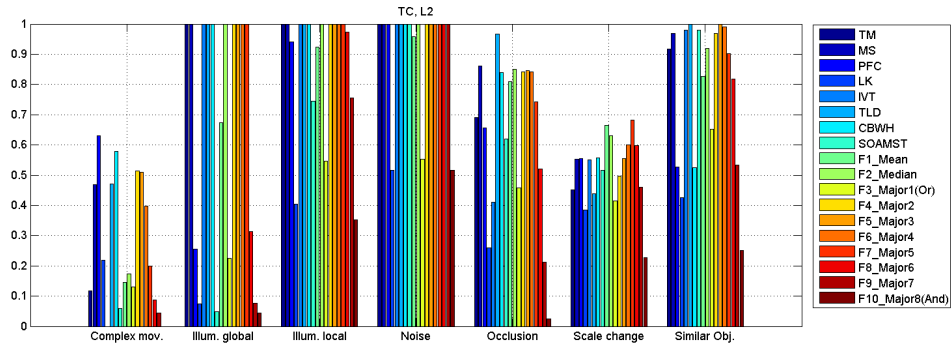


Figure C.30: Fusion TC result of L2

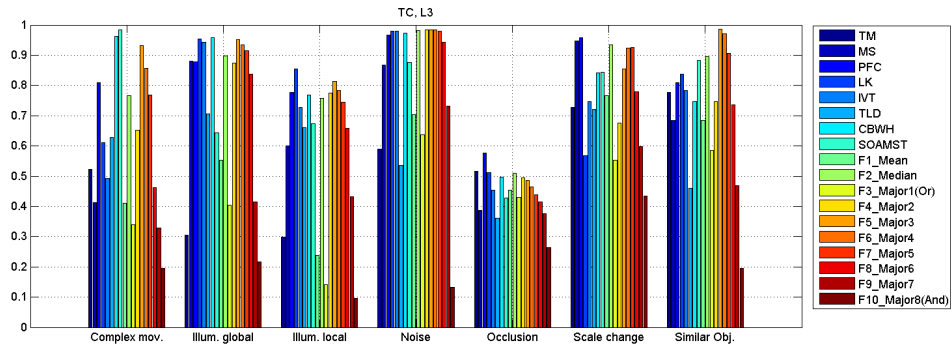


Figure C.31: Fusion TC result of L3

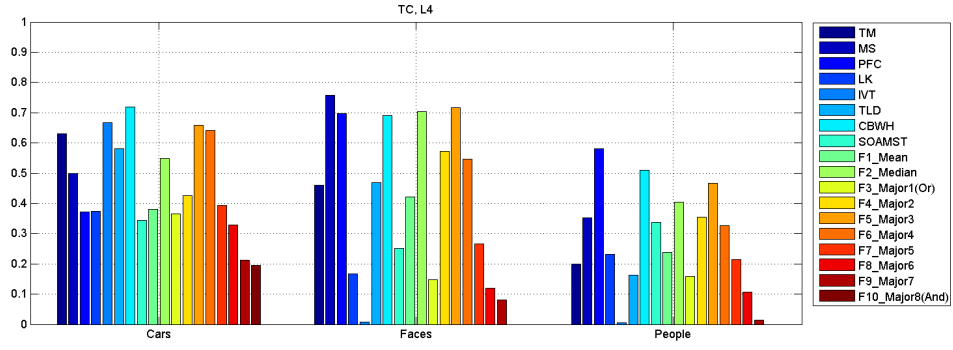


Figure C.32: Fusion TC result of L4

## C.8 Combined Tracking Performance Score (CoTPS)

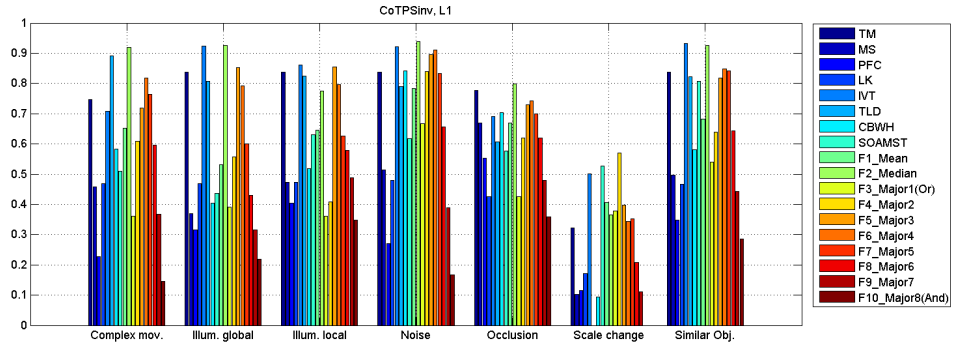


Figure C.33: Fusion CoTPSinv result of L1

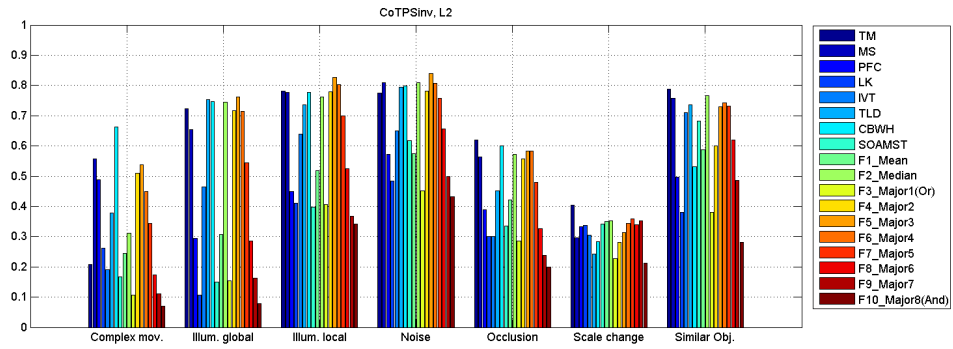


Figure C.34: Fusion CoTPSinv result of L2



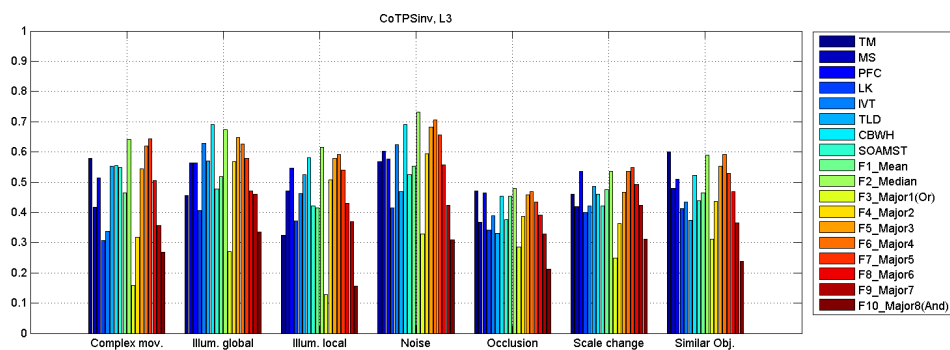


Figure C.35: Fusion CoTPSinv result of L3

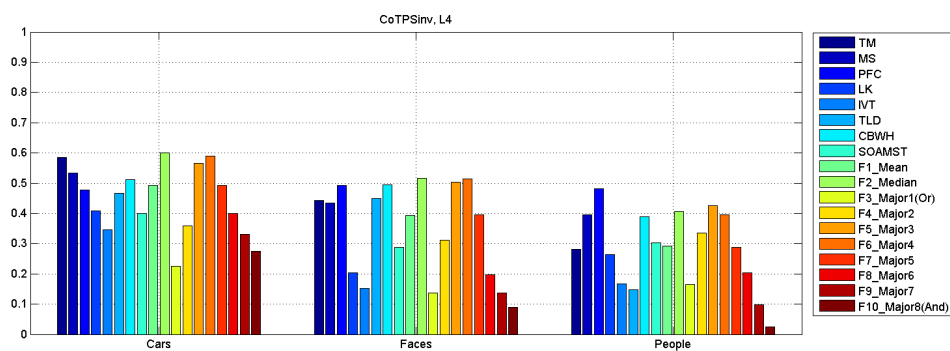


Figure C.36: Fusion CoTPSinv result of L4



## Appendix D

# Trackers and fusion results: Comparative tables

This appendix presents the obtained fusion results with tables. Four figures (one for each dataset category) are presented for each one of the nine metrics.

### D.1 Individual and fusions global scores

The table contained in this section presents the individual global scores for each metric and each tracker. These global scores are obtained as follows: For each metric and each tracker, all the dataset subcategories results are summed. It is, 7 for the L1 category, 7 for the L2 category, 7 for the L3 category and 3 for the L4 category. The result is then normalized between 0 and 1, dividing the result by 24 (7+7+7+3).

Tables D.1 and D.2 presents the global scores for the individual and fusion trackers respectively.

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
SFDA	0,481	0,372	0,319	0,246	0,424	0,429	0,429	0,348
ATA	0,485	0,375	0,321	0,249	0,427	0,432	0,432	0,351
ATEinv	0,571	0,465	0,626	0,541	0,656	0,676	0,526	0,712
AUCinv	0,518	0,430	0,366	0,276	0,484	0,489	0,494	0,386
PixelOv	0,523	0,434	0,370	0,279	0,488	0,492	0,499	0,389
CTM	0,485	0,375	0,321	0,249	0,427	0,432	0,432	0,351
CTD	0,167	0,175	0,192	0,117	0,095	0,126	0,177	0,153
TC	0,688	0,684	0,615	0,603	0,703	0,698	0,731	0,653
CoTPSinv	0,594	0,508	0,434	0,365	0,532	0,550	0,561	0,458

Table D.1: Individual global scores

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
SFDA	0,349	0,507	0,291	0,462	0,525	0,511	0,423	0,308	0,196	0,078
ATA	0,352	0,511	0,293	0,465	0,529	0,515	0,427	0,311	0,198	0,079
ATEinv	0,540	0,676	0,303	0,499	0,630	0,728	0,788	0,820	0,868	0,919
AUCinv	0,406	0,597	0,306	0,513	0,606	0,594	0,489	0,350	0,216	0,082
PixelOv	0,410	0,601	0,312	0,517	0,610	0,599	0,493	0,353	0,219	0,083
CTM	0,352	0,511	0,293	0,465	0,529	0,515	0,427	0,311	0,198	0,079
CTD	0,195	0,180	0,171	0,158	0,150	0,158	0,174	0,163	0,138	0,088
TC	0,637	0,781	0,497	0,770	0,847	0,813	0,715	0,581	0,445	0,192
CoTPSinv	0,496	0,656	0,308	0,527	0,637	0,639	0,572	0,449	0,338	0,223

Table D.2: Fusion global scores

## D.2 Difference global scores

This section presents the differences between the global scores obtained for each fusion and the global scores obtained for each individual tracker algorithm.

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-13,20	-2,33	3,01	10,26	-7,48	-8,01	-7,99	0,09
F2_Median	2,57	13,45	18,79	26,03	8,29	7,77	7,78	15,86
F3_Major1	-19,02	-8,14	-2,80	4,44	-13,29	-13,82	-13,81	-5,73
F4_Major2	-1,93	8,94	14,29	21,53	3,79	3,26	3,28	11,36
F5_Major3	<b>4,41</b>	<b>15,28</b>	<b>20,63</b>	<b>27,87</b>	<b>10,13</b>	<b>9,60</b>	<b>9,62</b>	<b>17,70</b>
F6_Major4	3,02	13,90	19,24	26,48	8,75	8,22	8,23	16,31
F7_Major5	-5,77	5,10	10,45	17,69	-0,05	-0,58	-0,56	7,52
F8_Major6	-17,27	-6,40	-1,05	6,19	-11,55	-12,08	-12,06	-3,98
F9_Major7	-28,50	-17,63	-12,28	-5,04	-22,78	-23,31	-23,29	-15,21
F10_Major8	-40,30	-29,43	-24,08	-16,84	-34,58	-35,11	-35,09	-27,01

Table D.3: Difference (percentaje) SFDA global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-13,31	-2,33	3,03	10,31	-7,56	-8,06	-8,05	0,07
F2_Median	2,60	13,58	18,94	26,22	8,35	7,85	7,86	15,98
F3_Major1	-19,16	-8,18	-2,82	4,46	-13,41	-13,91	-13,90	-5,78
F4_Major2	-1,94	9,04	14,40	21,67	3,80	3,31	3,31	11,43
F5_Major3	<b>4,45</b>	<b>15,44</b>	<b>20,79</b>	<b>28,07</b>	<b>10,20</b>	<b>9,70</b>	<b>9,71</b>	<b>17,83</b>
F6_Major4	3,05	14,04	19,39	26,67	8,80	8,30	8,31	16,43
F7_Major5	-5,81	5,17	10,53	17,80	-0,07	-0,56	-0,56	7,56
F8_Major6	-17,40	-6,42	-1,06	6,21	-11,65	-12,15	-12,15	-4,02
F9_Major7	-28,72	-17,74	-12,38	-5,11	-22,98	-23,47	-23,47	-15,35
F10_Major8	-40,62	-29,64	-24,28	-17,01	-34,87	-35,37	-35,37	-27,25

Table D.4: Difference (percentaje) ATA global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-3,12	7,58	-8,55	-0,05	-11,56	-13,59	1,47	-17,17
F2_Median	10,44	21,13	5,00	13,50	1,99	-0,04	15,02	-3,62
F3_Major1	-26,89	-16,20	-32,33	-23,83	-35,33	-37,37	-22,31	-40,95
F4_Major2	-7,22	3,47	-12,66	-4,15	-15,66	-17,69	-2,63	-21,27
F5_Major3	5,83	16,52	0,39	8,89	-2,61	-4,65	10,41	-8,23
F6_Major4	15,68	26,37	10,24	18,75	7,24	5,21	20,27	1,63
F7_Major5	21,62	32,31	16,18	24,69	13,18	11,15	26,21	<b>7,57</b>
F8_Major6	24,81	35,50	19,37	27,87	16,37	14,33	29,39	10,75
F9_Major7	29,62	40,32	24,19	32,69	21,18	19,15	34,21	15,57
F10_Major8	<b>34,78</b>	<b>45,47</b>	<b>29,34</b>	<b>37,84</b>	<b>26,33</b>	<b>24,30</b>	<b>39,36</b>	<b>20,72</b>

Table D.5: Difference (percentaje) ATEinv global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-11,16	-2,32	4,00	13,04	-7,76	-8,22	-8,79	2,05
F2_Median	7,90	16,74	23,06	32,10	11,30	10,84	10,27	21,11
F3_Major1	-21,17	-12,33	-6,01	3,03	-17,77	-18,23	-18,80	-7,96
F4_Major2	-0,54	8,29	14,61	23,66	2,85	2,39	1,83	12,66
F5_Major3	<b>8,79</b>	<b>17,63</b>	<b>23,94</b>	<b>32,99</b>	<b>12,19</b>	<b>11,72</b>	<b>11,16</b>	<b>21,99</b>
F6_Major4	7,62	16,45	22,77	31,82	11,02	10,55	9,99	20,82
F7_Major5	-2,85	5,98	12,30	21,35	0,55	0,08	-0,48	10,35
F8_Major6	-16,84	-8,01	-1,69	7,35	-13,45	-13,91	-14,47	-3,64
F9_Major7	-30,15	-21,31	-15,00	-5,95	-26,75	-27,21	-27,78	-16,94
F10_Major8	-43,62	-34,78	-28,46	-19,42	-40,22	-40,68	-41,25	-30,41

Table D.6: Difference (percentaje) AUCinv global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-11,23	-2,33	3,98	13,15	-7,77	-8,21	-8,85	2,08
F2_Median	7,83	16,73	23,05	32,21	11,29	10,85	10,21	21,15
F3_Major1	-21,10	-12,21	-5,89	3,28	-17,64	-18,08	-18,72	-7,79
F4_Major2	-0,54	8,36	14,68	23,84	2,92	2,49	1,85	12,78
F5_Major3	<b>8,79</b>	<b>17,68</b>	<b>24,00</b>	<b>33,16</b>	<b>12,25</b>	<b>11,81</b>	<b>11,17</b>	<b>22,10</b>
F6_Major4	7,61	16,51	22,83	31,99	11,07	10,63	9,99	20,92
F7_Major5	-2,91	5,99	12,31	21,47	0,55	0,11	-0,53	10,41
F8_Major6	-16,98	-8,09	-1,77	7,39	-13,53	-13,96	-14,60	-3,67
F9_Major7	-30,36	-21,47	-15,15	-5,99	-26,91	-27,34	-27,98	-17,05
F10_Major8	-43,91	-35,02	-28,70	-19,54	-40,45	-40,89	-41,53	-30,60

Table D.7: Difference (percentaje) PixelOV global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-13,31	-2,33	3,03	10,31	-7,56	-8,06	-8,05	0,07
F2_Median	2,60	13,58	18,94	26,22	8,35	7,85	7,86	15,98
F3_Major1	-19,16	-8,18	-2,82	4,46	-13,41	-13,91	-13,90	-5,78
F4_Major2	-1,94	9,04	14,40	21,67	3,80	3,31	3,31	11,43
F5_Major3	<b>4,45</b>	<b>15,44</b>	<b>20,79</b>	<b>28,07</b>	<b>10,20</b>	<b>9,70</b>	<b>9,71</b>	<b>17,83</b>
F6_Major4	3,05	14,04	19,39	26,67	8,80	8,30	8,31	16,43
F7_Major5	-5,81	5,17	10,53	17,80	-0,07	-0,56	-0,56	7,56
F8_Major6	-17,40	-6,42	-1,06	6,21	-11,65	-12,15	-12,15	-4,02
F9_Major7	-28,72	-17,74	-12,38	-5,11	-22,98	-23,47	-23,47	-15,35
F10_Major8	-40,62	-29,64	-24,28	-17,01	-34,87	-35,37	-35,37	-27,25

Table D.8: Difference (percentaje) CTM global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	<b>2,87</b>	<b>2,03</b>	<b>0,34</b>	<b>7,81</b>	<b>9,99</b>	<b>6,90</b>	<b>1,85</b>	<b>4,21</b>
F2_Median	1,33	0,49	-1,20	6,26	8,45	5,35	0,31	2,66
F3_Major1	0,38	-0,46	-2,15	5,32	7,50	4,41	-0,64	1,72
F4_Major2	-0,89	-1,73	-3,42	4,04	6,23	3,13	-1,91	0,44
F5_Major3	-1,66	-2,50	-4,19	3,28	5,46	2,37	-2,68	-0,32
F6_Major4	-0,86	-1,70	-3,39	4,07	6,26	3,17	-1,88	0,47
F7_Major5	0,78	-0,06	-1,75	5,71	7,90	4,81	-0,24	2,11
F8_Major6	-0,38	-1,23	-2,92	4,55	6,74	3,64	-1,41	0,95
F9_Major7	-2,91	-3,75	-5,44	2,03	4,21	1,12	-3,93	-1,57
F10_Major8	-7,91	-8,75	-10,44	-2,98	-0,79	-3,89	-8,93	-6,58

Table D.9: Difference (percentaje) CTD global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-5,09	-4,72	2,19	3,44	-6,62	-6,12	-9,37	-1,58
F2_Median	9,30	9,67	16,58	17,83	7,76	8,26	5,01	12,80
F3_Major1	-19,07	-18,71	-11,79	-10,54	-20,61	-20,11	-23,36	-15,57
F4_Major2	8,26	8,63	15,54	16,79	6,72	7,22	3,98	11,76
F5_Major3	<b>15,95</b>	<b>16,32</b>	<b>23,23</b>	<b>24,48</b>	<b>14,41</b>	<b>14,91</b>	<b>11,67</b>	<b>19,45</b>
F6_Major4	12,49	12,85	19,76	21,01	10,95	11,45	8,20	15,99
F7_Major5	2,76	3,12	10,03	11,28	1,22	1,72	-1,53	6,26
F8_Major6	-10,70	-10,34	-3,43	-2,18	-12,24	-11,74	-14,99	-7,20
F9_Major7	-24,33	-23,96	-17,05	-15,80	-25,87	-25,37	-28,61	-20,83
F10_Major8	-49,59	-49,23	-42,32	-41,06	-51,13	-50,63	-53,88	-46,09

Table D.10: Difference (percentaje) TC global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-9,86	-1,19	6,17	13,06	-3,59	-5,41	-6,54	3,77
F2_Median	<b>6,19</b>	<b>14,86</b>	<b>22,21</b>	<b>29,11</b>	<b>12,45</b>	<b>10,64</b>	<b>9,50</b>	<b>19,81</b>
F3_Major1	-28,65	-19,98	-12,63	-5,73	-22,39	-24,20	-25,34	-15,03
F4_Major2	-6,75	1,92	9,27	16,17	-0,49	-2,30	-3,44	6,87
F5_Major3	4,23	12,90	20,25	27,15	10,49	8,68	7,54	17,85
F6_Major4	4,42	13,09	20,45	27,34	10,69	8,87	7,74	18,05
F7_Major5	-2,20	6,47	13,83	20,72	4,07	2,25	1,12	11,43
F8_Major6	-14,56	-5,89	1,46	8,36	-8,30	-10,12	-11,25	-0,94
F9_Major7	-25,67	-17,00	-9,64	-2,75	-19,40	-21,22	-22,35	-12,04
F10_Major8	-37,12	-28,45	-21,09	-14,20	-30,85	-32,67	-33,80	-23,49

Table D.11: Difference (percentaje) CoTPS global score between fusion trackers and individual algorithms trackers

### D.3 Percentual difference global scores

This section presents the percentual differences between the global scores obtained for each fusion and the global scores obtained for each individual tracker algorithm. This result is obtained as presented in ecuation D.1.

$$Percentual\ difference = \frac{fusion\ score - individual\ score}{individual\ score} \quad (D.1)$$

Below are the different tables.

APPENDIX D. TRACKERS AND FUSION RESULTS: COMPARATIVE TABLES92

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-27,45	-6,26	9,46	41,63	-17,65	-18,67	-18,63	0,25
F2_Median	5,35	36,12	58,94	105,67	19,58	18,10	18,15	45,57
F3_Major1	-39,54	-21,88	-8,78	18,03	-31,38	-32,22	-32,19	-16,46
F4_Major2	-4,01	24,02	44,82	87,39	8,95	7,61	7,65	32,63
F5_Major3	<b>9,17</b>	<b>41,06</b>	<b>64,71</b>	<b>113,13</b>	<b>23,91</b>	<b>22,39</b>	<b>22,44</b>	<b>50,85</b>
F6_Major4	6,29	37,33	60,36	107,50	20,64	19,15	19,20	46,87
F7_Major5	-12,00	13,71	32,77	71,80	-0,11	-1,34	-1,30	21,60
F8_Major6	-35,91	-17,18	-3,30	25,13	-27,25	-28,15	-28,12	-11,43
F9_Major7	-59,26	-47,36	-38,53	-20,46	-53,76	-54,33	-54,31	-43,70
F10_Major8	-83,79	-79,06	-75,55	-68,36	-81,61	-81,83	-81,82	-77,61

Table D.12: Percentual difference (percentage) SFDA global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-27,45	-6,20	9,44	41,45	-17,70	-18,64	-18,63	0,19
F2_Median	5,37	36,23	58,94	105,44	19,53	18,16	18,17	45,51
F3_Major1	-39,52	-21,80	-8,77	17,92	-31,39	-32,17	-32,17	-16,47
F4_Major2	-4,01	24,10	44,79	87,15	8,89	7,65	7,66	32,56
F5_Major3	<b>9,19</b>	<b>41,17</b>	<b>64,71</b>	<b>112,89</b>	<b>23,87</b>	<b>22,45</b>	<b>22,46</b>	<b>50,79</b>
F6_Major4	6,30	37,43	60,35	107,25	20,59	19,21	19,22	46,80
F7_Major5	-11,99	13,79	32,76	71,60	-0,15	-1,30	-1,29	21,55
F8_Major6	-35,89	-17,11	-3,29	25,00	-27,27	-28,10	-28,10	-11,46
F9_Major7	-59,25	-47,32	-38,53	-20,55	-53,77	-54,30	-54,30	-43,73
F10_Major8	-83,79	-79,05	-75,55	-68,40	-81,61	-81,82	-81,82	-77,62

Table D.13: Percentual difference (percentaje) ATA global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-5,45	16,31	-13,67	-0,10	-17,62	-20,10	2,79	-24,11
F2_Median	18,26	45,48	7,98	24,96	3,04	-0,06	28,57	-5,08
F3_Major1	-47,06	-34,87	-51,66	-44,06	-53,87	-55,26	-42,44	-57,51
F4_Major2	-12,63	7,48	-20,22	-7,68	-23,88	-26,16	-5,01	-29,88
F5_Major3	10,20	35,56	0,62	16,44	-3,99	-6,87	19,81	-11,55
F6_Major4	27,44	56,77	16,37	34,66	11,04	7,70	38,55	2,29
F7_Major5	37,84	69,56	25,86	45,64	20,09	16,49	49,86	10,63
F8_Major6	43,41	76,42	30,95	51,53	24,95	21,20	55,92	15,10
F9_Major7	51,84	86,78	38,64	60,44	32,29	28,32	65,08	21,86
F10_Major8	<b>60,85</b>	<b>97,87</b>	<b>46,88</b>	<b>69,96</b>	<b>40,15</b>	<b>35,94</b>	<b>74,88</b>	<b>29,10</b>

Table D.14: Percentual difference (percentaje) ATEinv global score between fusion trackers and individual algorithms trackers



APPENDIX D. TRACKERS AND FUSION RESULTS: COMPARATIVE TABLES93

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-21,54	-5,40	10,91	47,26	-16,03	-16,82	-17,77	5,31
F2_Median	15,26	38,96	62,92	116,31	23,35	22,18	20,78	54,69
F3_Major1	-40,86	-28,70	-16,41	10,99	-36,71	-37,31	-38,03	-20,63
F4_Major2	-1,05	19,31	39,87	85,71	5,90	4,90	3,70	32,81
F5_Major3	<b>16,97</b>	<b>41,03</b>	<b>65,34</b>	<b>119,51</b>	<b>25,18</b>	<b>24,00</b>	<b>22,57</b>	<b>56,99</b>
F6_Major4	14,71	38,30	62,14	115,27	22,76	21,60	20,21	53,95
F7_Major5	-5,51	13,93	33,57	77,34	1,13	0,17	-0,98	26,82
F8_Major6	-32,52	-18,64	-4,61	26,64	-27,78	-28,46	-29,28	-9,43
F9_Major7	-58,20	-49,61	-40,92	-21,56	-55,27	-55,69	-56,20	-43,90
F10_Major8	-84,20	-80,96	-77,67	-70,36	-83,10	-83,26	-83,45	-78,80

Table D.15: Percentual difference (percentaje) AUCinv global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-21,49	-5,39	10,76	47,16	-15,93	-16,67	-17,74	5,35
F2_Median	14,99	38,58	62,23	115,54	23,14	22,05	20,48	54,31
F3_Major1	-40,38	-28,15	-15,89	11,75	-36,16	-36,72	-37,54	-20,00
F4_Major2	-1,03	19,28	39,63	85,52	5,99	5,05	3,70	32,82
F5_Major3	<b>16,82</b>	<b>40,79</b>	<b>64,81</b>	<b>118,98</b>	<b>25,10</b>	<b>23,99</b>	<b>22,40</b>	<b>56,76</b>
F6_Major4	14,56	38,07	61,63	114,75	22,68	21,60	20,04	53,74
F7_Major5	-5,57	13,81	33,23	77,01	1,12	0,23	-1,05	26,72
F8_Major6	-32,50	-18,66	-4,78	26,52	-27,72	-28,36	-29,28	-9,42
F9_Major7	-58,11	-49,51	-40,90	-21,48	-55,14	-55,54	-56,11	-43,78
F10_Major8	-84,04	-80,76	-77,48	-70,08	-82,91	-83,06	-83,28	-78,58

Table D.16: Percentual difference (percentaje) PixelOV global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-27,45	-6,20	9,44	41,45	-17,70	-18,64	-18,63	0,19
F2_Median	5,37	36,23	58,94	105,44	19,53	18,16	18,17	45,51
F3_Major1	-39,52	-21,80	-8,77	17,92	-31,39	-32,17	-32,17	-16,47
F4_Major2	-4,01	24,10	44,79	87,15	8,89	7,65	7,66	32,56
F5_Major3	<b>9,19</b>	<b>41,17</b>	<b>64,71</b>	<b>112,89</b>	<b>23,87</b>	<b>22,45</b>	<b>22,46</b>	<b>50,79</b>
F6_Major4	6,30	37,43	60,35	107,25	20,59	19,21	19,22	46,80
F7_Major5	-11,99	13,79	32,76	71,60	-0,15	-1,30	-1,29	21,55
F8_Major6	-35,89	-17,11	-3,29	25,00	-27,27	-28,10	-28,10	-11,46
F9_Major7	-59,25	-47,32	-38,53	-20,55	-53,77	-54,30	-54,30	-43,73
F10_Major8	-83,79	-79,05	-75,55	-68,40	-81,61	-81,82	-81,82	-77,62

Table D.17: Percentual difference (percentaje) CTM global score between fusion trackers and individual algorithms trackers

APPENDIX D. TRACKERS AND FUSION RESULTS: COMPARATIVE TABLES94

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	<b>17,23</b>	<b>11,60</b>	<b>1,77</b>	<b>66,51</b>	<b>104,66</b>	<b>54,55</b>	<b>10,45</b>	<b>27,42</b>
F2_Median	7,98	2,79	-6,26	53,37	88,50	42,35	1,74	17,36
F3_Major1	2,31	-2,61	-11,18	45,31	78,60	34,87	-3,61	11,20
F4_Major2	-5,35	-9,90	-17,83	34,44	65,23	24,78	-10,82	2,87
F5_Major3	-9,93	-14,26	-21,81	27,93	57,23	18,74	-15,14	-2,10
F6_Major4	-5,16	-9,71	-17,66	34,71	65,57	25,04	-10,64	3,09
F7_Major5	4,68	-0,35	-9,12	48,69	82,75	38,01	-1,37	13,78
F8_Major6	-2,30	-7,00	-15,19	38,76	70,55	28,80	-7,95	6,19
F9_Major7	-17,44	-21,40	-28,32	17,27	44,13	8,85	-22,21	-10,26
F10_Major8	-47,47	-49,99	-54,39	-25,38	-8,29	-30,74	-50,50	-42,90

Table D.18: Percentual difference (percentaje) CTD global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-7,39	-6,90	<b>3,57</b>	<b>5,71</b>	-9,42	-8,77	-12,82	-2,43
F2_Median	13,52	14,13	26,95	29,59	11,04	11,83	6,86	19,61
F3_Major1	-27,73	-27,34	-19,18	-17,50	-29,31	-28,80	-31,97	-23,85
F4_Major2	12,01	12,62	25,27	27,87	9,56	10,35	5,44	18,02
F5_Major3	<b>23,19</b>	<b>23,85</b>	<b>37,77</b>	<b>40,63</b>	<b>20,49</b>	<b>21,36</b>	<b>15,97</b>	<b>29,80</b>
F6_Major4	18,15	18,79	32,13	34,88	15,57	16,39	11,22	24,49
F7_Major5	4,01	4,56	16,31	18,73	1,73	2,46	-2,09	9,58
F8_Major6	-15,56	-15,11	-5,57	-3,61	-17,41	-16,82	-20,52	-11,04
F9_Major7	-35,37	-35,02	-27,72	-26,22	-36,78	-36,33	-39,16	-31,90
F10_Major8	-72,10	-71,95	-68,80	-68,15	-72,71	-72,52	-73,74	-70,60

Table D.19: Percentual difference (percentaje) TC global score between fusion trackers and individual algorithms trackers

	TM	MS	PFC	LK	IVT	TLD	CBWH	SOAMST
F1_Mean	-16,59	-2,34	14,20	35,77	-6,75	-9,84	-11,66	8,22
F2_Median	<b>10,41</b>	<b>29,26</b>	<b>51,16</b>	<b>79,70</b>	<b>23,42</b>	<b>19,34</b>	<b>16,93</b>	<b>43,24</b>
F3_Major1	-48,20	-39,36	-29,08	-15,69	-42,10	-44,01	-45,14	-32,80
F4_Major2	-11,36	3,78	21,36	44,27	-0,91	-4,19	-6,12	15,00
F5_Major3	7,11	25,40	46,64	74,34	19,74	15,77	13,44	38,96
F6_Major4	7,43	25,78	47,09	74,86	20,10	16,13	13,78	39,38
F7_Major5	-3,70	12,75	31,84	56,74	7,65	4,09	1,99	24,94
F8_Major6	-24,50	-11,61	3,36	22,88	-15,60	-18,39	-20,04	-2,05
F9_Major7	-43,18	-33,48	-22,21	-7,52	-36,48	-38,59	-39,83	-26,29
F10_Major8	-62,44	-56,03	-48,58	-38,87	-58,02	-59,40	-60,22	-51,27

Table D.20: Percentual difference (percentaje) CoTPS global score between fusion trackers and individual algorithms trackers

## D.4 Difference and percentual difference global scores compared with the best individual tracker

This section presents the differences and percentual differences (see section D.3) between the global scores obtained for each fusion and the best one of global scores obtained for all the individual tracker algorithm.

	SFDA	ATA	ATEinv	AUCinv	PixelOv	CTM	TC	CoTPSinv
F1_Mean	-13,20	-13,31	-17,17	-11,16	-11,23	-13,31	-9,37	-9,86
F2_Median	<b>2,57</b>	2,60	-3,62	7,90	7,83	2,60	5,01	<b>6,19</b>
F3_Major1	-19,02	-19,16	-40,95	-21,17	-21,10	-19,16	-23,36	-28,65
F4_Major2	-1,93	-1,94	-21,27	-0,54	-0,54	-1,94	3,98	-6,75
F5_Major3	<b>4,41</b>	<b>4,45</b>	-8,23	<b>8,79</b>	<b>8,79</b>	<b>4,45</b>	<b>11,67</b>	4,23
F6_Major4	3,02	3,05	1,63	7,62	7,61	3,05	8,20	4,42
F7_Major5	-5,77	-5,81	7,57	-2,85	-2,91	-5,81	-1,53	-2,20
F8_Major6	-17,27	-17,40	10,75	-16,84	-16,98	-17,40	-14,99	-14,57
F9_Major7	-28,50	-28,72	15,57	-30,15	-30,36	-28,72	-28,61	-25,67
F10_Major8	-40,30	-40,62	<b>20,72</b>	-43,62	-43,91	-40,62	-53,88	-37,119

Table D.21: Difference (percentaje) global score between fusion trackers and best individual algorithms trackers

	SFDA	ATA	ATEinv	AUCinv	PixelOv	CTM	TC	CoTPSinv
F1_Mean	-27,45	-27,45	-24,11	-21,54	-21,49	-27,45	-12,82	-16,59
F2_Median	5,35	5,37	-5,08	15,26	14,99	5,37	6,86	<b>10,41</b>
F3_Major1	-39,54	-39,52	-57,51	-40,86	-40,38	-39,52	-31,97	-48,20
F4_Major2	-4,01	-4,01	-29,88	-1,05	-1,03	-4,01	5,44	-11,36
F5_Major3	<b>9,17</b>	<b>9,19</b>	-11,55	<b>16,97</b>	<b>16,82</b>	<b>9,19</b>	<b>15,97</b>	7,11
F6_Major4	6,29	6,30	2,29	14,71	14,56	6,30	11,22	7,43
F7_Major5	-12,00	-11,99	10,63	-5,51	-5,57	-11,99	-2,09	-3,70
F8_Major6	-35,91	-35,89	15,10	-32,52	-32,50	-35,89	-20,52	-24,50
F9_Major7	-59,26	-59,25	21,86	-58,20	-58,11	-59,25	-39,16	-43,18
F10_Major8	-83,79	-83,79	<b>29,10</b>	-84,20	-84,04	-83,79	-73,74	-62,44

Table D.22: Percentual difference (percentaje) global score between fusion trackers and best individual algorithms trackers